UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

DANIEL NEHME MÜLLER

COMFALA - Modelo Computacional do Processo de Compreensão da Fala

Tese apresentada como requisito parcial
para a obtenção do grau de
Doutor em Ciência da Computação

Prof. Dr. Philippe Olivier Alexandre Navaux
Orientador

Porto Alegre, março de 2006

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Müller, Daniel Nehme

COMFALA - Modelo Computacional do Processo de Compreensão da Fala / Daniel Nehme Müller. – Porto Alegre: PPGC da UFRGS, 2006.

130 f.: il.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2006. Orientador: Philippe Olivier Alexandre Navaux.

1. Compreensão da Linguagem Falada. 2. Reconhecimento de Voz. 3. Processamento de Linguagem Natural. I. Navaux, Philippe Olivier Alexandre. II. Título.

“Assim é que, não podendo os órgãos auditivos perceber as vibrações sonoras aquém das 32 e além das 32.768 emitidas pelos corpos num segundo, ficamos privados de um número de conhecimentos e verdades que superam, incomparavelmente, os conhecimentos e verdades relativos ao mundo exterior que possuímos e hão de possuir todas as gerações de perscrutadores que venham a habitar a superfície da Terra.

De fato, se dispusessemos de outros sentidos capazes de perceber certas vibrações, ou os nossos sentidos tivessem a faculdade de adaptar-se a essas variedades de vibrações, o órgão, por exemplo, da audição poderia perceber sons tão surpreendentes e maravilhosos que o deixariam como em uma espécie de êxtase contínuo. E dado que só pudesse perceber esses sons, então outra singularidade: não perceberia nenhum dos sons compreendidos entre 32 e 32.768 vibrações emitidas pelos corpos sonoros e perceptíveis. Reinaria dentro e em volta de nós um silêncio profundo; nada do que ouvimos agora poderíamos ouvir, nem mesmo a voz do nosso semelhante, nem o grito das feras, nem o tiro do canhão, nem o retumbar do trovão ou estalar do raio. Mas, em compensação, perceberíamos a harmonia que existe no canto das aves, perceberíamos mesmo ao longe o zumbir dos insetos, mais forte ainda do que o canto dos pássaros agora.

E se nossos ouvidos se adaptassem à percepção das vibrações sonoras entre 32.768 e 34 milhões, as maravilhas da audição ainda iriam mais longe e tocariam, até, às extremas fronteiras do sobrenatural. Porque, além da harmonia do canto dos pássaros, do zumbir ao longe dos insetos, além de todos os fenômenos sonoros, perceberíamos os fenômenos elétricos e magnéticos agora imperceptíveis, tais como os que se produzem ao despontar da aurora, ao surgir do Sol, ao aparecer dos astros ou desaparecer deles, ao relampejar, ao produzir-se a chispa elétrica, etc., etc. Perceberíamos todas essas vibrações tão clara e talvez ainda mais nitidamente do que percebemos as do nosso limitado mundo de percepções auditivas.”

— Pe. Roberto Landell de Moura

AGRADECIMENTOS

Agradeço a Deus, pelo desafio da vida.
Agradeço ao meu pai e minha mãe, pelo incentivo aos estudos.
Agradeço ao meu orientador, pela paciência e exigência.
Agradeço a minha esposa e minha filha, pelo apoio, carinho e compreensão.
Agradeço a todos os colegas que de alguma forma auxiliaram neste trabalho.
Agradeço a você que, lendo este trabalho no futuro, levará estas idéias adiante...

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ASR	Automatic Speech Recognition
ANN	Artificial Neural Network
CART	Classification and Regression Trees
COMFALA	Modelo Computacional de Compreensão de Fala
CRF	Chunk Relation Finder
DM	Dialogue Model
EEG	Eletroencefalografia
ELAN	Early Left-Anterior Negativity
ERP	Event-Related brain Potential
F0	Freqüência Fundamental
F1	Primeiro Grupo de Formantes (Harmônicas)
fMRI	functional Magnetic Resonance Imaging
FST	Finit-State Transducer
GCI	Glottal Closure Instant
GSS	Graph Strutured Stack
HMM	Hidden Markov Models
IA	Inteligência Artificial
IM	Intonation Model
LAN	Left-Anterior Negativity
LFL	Linguistic Feature Labeler
LM	Language Model
LPC	Linear Predictive Coding
LPCo	Late Positive Component
MEG	Magnetoencefalografia
MNPAF	Modelo Neurocognitivo da Audição de Frases
MRI	Magnetic Resonance Imaging

NLP Natural Language Processing

PET Positron Imaging Tomography

PLN Processamento de Linguagem Natural

RAAM Recursive Auto-Associative Memory

RTN Recursive Transition Network

SARDNET Sequential Activation Retention and Decay Network

SCREEN Simbolic Connectionist Robust EnterprisE for Natural language

SLU Spoken Language Understanding

SOM Self-Organizing Map

SUM Speech Understanding Model

SRAAM Sequential Recursive Auto-Associative Memory

SRN Simple Recurrent Network

TBL Transformation-Based Learning

TDNN Time Delay Neural Network

TDRNN Time Delay Recursive Neural Network

TKM Temporal Kohonen Map

LISTA DE FIGURAS

RESUMO

Esta Tese apresenta a investigação de técnicas computacionais que permitam a simulação computacional da compreensão de frases faladas. Esta investigação é baseada em estudos neurocognitivos que descrevem o processamento do cérebro ao interpretar a audição de frases. A partir destes estudos, realiza-se a proposição do COMFALA, um modelo computacional para representação do processo de compreensão da fala. O COMFALA possui quatro módulos, correspondentes às fases do processamento cerebral: processamento do sinal de fala, análise sintática, análise semântica e avaliação das respostas das análises. Para validação do modelo são propostas implementações para cada módulo do COMFALA. A codificação do sinal se dá através das transformadas ondeletas (*wavelets transform*s), as quais permitem uma representação automática de padrões para sistemas conexionistas (redes neurais artificiais) responsáveis pela análise sintática e semântica da linguagem. Para a análise sintática foi adaptado um sistema conexionista de linguagem escrita. Por outro lado, o sistema conexionista de análise semântica realiza agrupamentos por características prosódicas e fonéticas do sinal. Ao final do processo, compara-se a saída sintática com a semântica, na busca de uma melhor interpretação da fala.

Palavras-chave: Compreensão da Linguagem Falada, Reconhecimento de Voz, Processamento de Linguagem Natural.

SUM - Speech Understanding Model

ABSTRACT

This thesis presents the investigation of computational technologies to allow the simulation of speech understanding. This investigation is based on neurocognitive researches to describe the brain's processes when interpreting heard sentences. Thus, we propose SUM, a Speech Understanding Model. SUM has four modules that correspond to the cerebral processing phases: speech signal processing, syntactic analysis, semantic analysis and responses evaluation of the analyses. To validate the model, implementations for each SUM module are proposed. This signal codification is based on wavelets transform and it makes possible an automatic representation of syntactic and semantic analysis carried out by connectionist systems (artificial neural network). The syntactic analysis connectionist system organizes syntactic structures and temporal identification of words in a sentence. On the other hand, the semantic analysis connectionist system shapes clusters of prosodic and semantic characteristics of the signal. In the end of the process, we compare the results of syntactic and semantic analysis to obtain a better interpretation of speech.

Keywords: Spoken Language Understanding, Automatic Speech Recognition, Natural Language Processing.

1 INTRODUÇÃO

A linguagem falada sem dúvida é a primeira forma de comunicação humana que tomamos contato. O bebê, desde a fase de gestação, escuta a voz da mãe falando-lhe. Ao nascer, ele já reconhece a mãe mais por suas características de fala do que pela visão, ainda em formação.

No decorrer de nosso desenvolvimento, a fala é determinante para a construção das relações com as outras pessoas. A forma de interação que constitui a linguagem falada também define simultaneamente nossa forma de organização do pensamento. Em síntese, a fala proporciona nossa comunicação primária com o mundo.

Além dos relacionamentos humanos e com a natureza, também criamos na vida moderna as relações com as máquinas. Para possibilitar a manipulação de equipamentos elétricos e eletrônicos, foram criadas interfaces de comunicação. A princípio, estas interfaces são baseadas em botões para a entrada de sinais do usuário e mostradores (*displays*) que permitem a saída da resposta do sistema.

Diariamente bilhões de pessoas utilizam dispositivos como os citados em seu cotidiano. Seja relógio, televisor, forno de microondas, telefone celular ou computador, todo o mundo eletrônico precisa de uma interface para que as pessoas possam controlar seus aparelhos. Anteriormente os dispositivos de manipulação móvel (botões, controles deslizantes) foram incorporados pelos desenhos e ícones em interfaces de sistemas computacionais. Atualmente tentam-se adaptar outros recursos de comunicação como a compreensão da fala e dos movimentos do corpo, no sentido de facilitar a manipulação dos dispositivos eletrônicos.

Todavia, a complexidade envolvida na comunicação humana vem há décadas intrigando os pesquisadores que buscam uma melhor forma de controle do computador e dos demais dispositivos eletrônicos. Foi constatado ao longo dos anos que não há como realizar uma modelagem simplista para o fenômeno da comunicação falada e visual. Disto decorrem modelagens igualmente complexas para se obter um resultado prático satisfatório e não serão abordadas na presente Tese.

Dada a complexidade das abordagens envolvidas, e apesar da importância que possui a comunicação visual para a compreensão, é necessário aqui limitar a análise à fala. Dentro desta abordagem, é enfocada aqui a modelagem da compreensão da linguagem falada, sem a preocupação de uma produção de linguagem em resposta à compreensão realizada. Os sistemas de produção de respostas têm sido extensamente desenvolvidos através do Processamento de Linguagem Natural (PLN) da linguagem escrita.

Por outro lado, um dos grandes enfrentamentos da Computação é a realização de uma adequada representação computacional da compreensão da fala. Existem inúmeros obstáculos práticos que originam esta dificuldade. Dentre estes, podem ser citados a representação do sinal sonoro da voz, seu seqüenciamento no tempo, a identificação dos elementos

fonéticos desta seqüência, a composição destes elementos fonéticos em função de um léxico conhecido, e a organização da seqüência de termos léxicos para a identificação de frases com sintaxe e semântica coerentes.

Cada uma das características citadas compreendem um campo de pesquisa em andamento na busca da representação computacional da compreensão da fala. Para melhor contextualização destas áreas de pesquisa, passa-se a seguir para uma breve análise da modelagem computacional da compreensão de fala que vem sendo constituída ao longo de duas décadas.

1.1 Questões e Hipótese de Pesquisa

O processamento computacional da compreensão da linguagem falada é um tópico de pesquisa que ainda deve ser muito explorado, por não haver uma tecnologia consolidada nesta área. Por isso podem ser encontrados diversos pontos de questionamento que levam a novos caminhos de pesquisa. Essas questões de pesquisa são apresentadas a seguir.

- É viável o desenvolvimento de um sistema integrado conexionista para compreensão de fala?

Freqüentemente encontram-se modelos que abordam apenas o reconhecimento de fala, apenas a análise da sintaxe da fala, ou apenas a semântica. São poucas as abordagens que salientam o modelo completo de compreensão. E menor é o número de modelos de diálogo, possuindo a compreensão e a produção da fala. Dentro do escopo da compreensão de fala, só foi constatada a abordagem do SCREEN, analisado na seção 2.1, como possuindo a análise sintática e semântica conexionistas. Observa-se aqui que a abordagem conexionista permite uma organização automática do modelo, uma vez que os sistemas envolvidos são treinados através de exemplos.

- A codificação da fala por transformadas de Fourier pode ser substituída por outra metodologia para melhor caracterização da linguagem falada contínua?

Sabe-se que a transformada de Fourier é adequada para sinais estacionários. Por outro lado, a voz tem por característica ser constituída por ondas não estacionárias. Em conseqüência disso, a análise temporal da onda é prejudicada, uma vez que Fourier dá espaços de tempo iguais a freqüências diferentes. Ondas com maior freqüência devem ter uma janela de análise menor, assim como ondas de menor freqüência devem ter janelas maiores de tempo.

- A análise da onda fundamental pelos algoritmos tradicionais é eficiente e suficiente para a obtenção dos dados prosódicos?

Busca-se a obtenção de características do sinal que permitam uma representação da prosódia através de transformadas aplicadas diretamente ao sinal. Atualmente são realizadas seqüências refinadas de algoritmos para estimar a onda fundamental da fala.

- Modelos conexionistas podem ser utilizados como alternativa ao HMM (*Hidden Markov Model* - Modelo Oculto de Markov) para modelagem da linguagem?

Os métodos usados normalmente com o HMM necessitam da modelagem manual das cadeias de probabilidade de ocorrência temporal de fonemas, palavras ou frases. Procuram-se alternativas automatizadas nos modelos conexionistas, uma vez que estes são treinados através de exemplos.

- Modelos conexionistas podem ser uma alternativa viável à árvore de análise da linguagem?

De forma semelhante à questão anterior, buscam-se modelos de construção automatizada. As árvores de análise da linguagem atualmente são previamente definidas manualmente.

- Os dados prosódicos podem ser inseridos diretamente nos processos de análise sintática e semântica?

Diversas pesquisas constataram que a expressão da fala pode auxiliar na resolução de ambigüidades sintáticas e semânticas. Por outro lado, estas pesquisas utilizaram os dados prosódicos como uma informação a mais utilizada em árvores de análise da linguagem. Em outras palavras, também são resultado de uma modelagem manual.

- Como ocorre o processo natural da compreensão da fala?

A modelagem computacional da compreensão da fala deve refletir primeiramente a realidade constatada em pesquisas neuropsicológicas. Isso porque todo modelo é uma representação do mundo real. Se a realidade em questão é a fala, a realidade do ser humano deve ser avaliada para tal concepção.

Com base nos questionamentos realizados, foi constituída uma hipótese que norteou o desenvolvimento da presente Tese:

- *Deve ser possível a constituição de um modelo computacional que tenha por base pesquisas da compreensão do processamento natural da compreensão da fala. Este modelo deve permitir uma representação temporal do sinal da fala e de sua prosódia. Os dados obtidos na representação do sinal devem servir de base para as análises sintática e semântica da linguagem. Estas análises devem ser construídas automaticamente, a partir de exemplos de fala.*

1.2 Metodologia

A investigação da hipótese de pesquisa foi iniciada com um levantamento bibliográfico em livros e periódicos acerca do estado-da-arte dos temas envolvidos nas questões de pesquisa. Após o levantamento, foram selecionados livros e artigos segundo o critério da viabilização de um modelo de compreensão de fala. Neste sentido, os temas abordados deveriam responder às questões de pesquisa.

Após organizado e compilado o material bibliográfico, passou-se ao registro escrito e aos testes de protótipo. Foram pesquisados protótipos existentes e a possibilidade de uso, aproveitamento ou aperfeiçoamento de sistemas ou ferramentas existentes. Além dos sistemas pesquisados, foram desenvolvidos novos sistemas para viabilização do modelo proposto.

Através dos protótipos foram procedidos testes de verificação de viabilidade dos sistemas, métodos e técnicas pesquisadas. Disso resultou uma seleção de ferramentas para a Tese. Com base nestes primeiros resultados, foram efetivados estudos em profundidade e desenvolvidos novos testes, agora para validação do modelo.

Todos os passos realizados foram registrados no texto da presente Tese, de forma a permitir a continuidade desta pesquisa. Tem-se consciência que, fruto deste trabalho, advirão diversos projetos de pesquisa derivados.

1.3 Organização da Tese

O texto da presente Tese foi organizado de forma a dar sustentação teórica e prática à hipótese de pesquisa. O início é dado no capítulo 2 com a contextualização acerca de sistemas de compreensão de fala e correlatos. Desta forma, o capítulo 3 apresenta o modelo natural que irá nortear o processamento computacional. No capítulo 4 será apresentada a forma de solução encontrada para codificação do sinal da fala. A abordagem utilizada para a análise sintática e semântica será tratada no capítulo 5. O modelo computacional que sintetiza todo o trabalho desenvolvido está apresentado no capítulo 6. Por fim serão apresentadas simulações computacionais destes processos no capítulo 7 e as conclusões desta Tese no capítulo 8.

2 SISTEMAS COMPUTACIONAIS PARA COMPREENSÃO DA FALA

Os primeiros sistemas que buscaram a compreensão da fala remontam aos anos 1980, com o SUMMIT-TINA, desenvolvido no Instituto Tecnológico de Massachussets (MIT), o SPICOS II, do consórcio SIEMENS-PHILIPS-IPO, e SUNDIAL, sistema participante do projeto ESPRIT (LARSEN et al., 1992). O sistema SUMMIT realizava reconhecimento de fala independente do usuário, ao qual era acoplado o TINA, que fazia a interpretação da fala usando uma gramática livre de contexto com transições probabilísticas. O SPICOS II foi implementado através de modelagem por cadeias de Markov para o reconhecimento de fala e redes semânticas para a representação da linguagem. De forma semelhante, o sistema SUNDIAL também tinha o reconhecimento de fala realizado com HMMs (*Hidden Markov Models* - Modelos Ocultos de Markov - veja seção 5.3) e a interpretação da linguagem por redes semânticas e quadros (*frames*).

2.1 SCREEN

Os anos 1990 foram marcados pelo uso de sistemas com forte enfoque probabilístico e a inserção de sistemas conexionistas como uma das alternativas de implementação dos modelos de compreensão de fala. Um sistema que representa esta nova abordagem é o sistema SCREEN, de Wermter e Weber (WEBER; WERMTER, 1996; WERMTER; WEBER, 1997) e Wermter e Löchel (WERMTER; LÖCHEL, 1996). O sistema SCREEN (*Symbolic Connectionist Robust EnterprisE for Natural language*) é constituído de seis partes, conforme a figura 2.1: construção da seqüência de fala, avaliação da fala, categoria, correção, casos e diálogo. Cada parte é formada por módulos conexionistas e/ou simbólicos que geram hipóteses de frases para a parte seguinte, num nível crescente de abstração da linguagem.

A construção da **seqüência da fala** é produzida a partir do resultado de um reconhecedor HMM. Nesta parte são geradas diversas hipóteses de frases através da combinação das palavras contidas no léxico do sistema. A **parte de avaliação** da fala realiza uma estimativa de erros provocados pela fala. As hipóteses de palavras oriundas do reconhecedor são avaliadas sintática e semanticamente para a constituição de frases através de redes neurais recorrentes simples (SRN - *Simple Recurrent Network*). As frases de melhor avaliação vão para a **parte de categoria**, onde são resolvidas possíveis ambigüidades resultantes da classificação sintática e semântica. Para tanto, é utilizada uma SRN para analisar a sintaxe e outra para a semântica. Os resultados destas análises são utilizados posteriormente na parte de correção de fala e são também passados a outras duas SRNs ainda na parte de categoria, onde farão uma classificação em termos de categorias abstra-

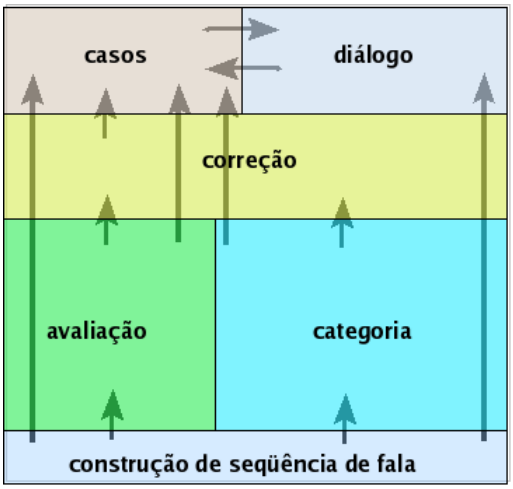

Figura 2.1: Estrutura do Sistema SCREEN.

tas sintáticas (grupo verbal, grupo de substantivo, etc.) e categorias abstratas semânticas (ação, agente da ação, etc.).

Já a **parte de correção** identifica pausas, interjeições e corrige as palavras e a estrutura da frase. A correção de palavras é feita com base na comparação das classes sintáticas e semânticas básicas entre duas palavras. Para isso utilizam-se redes neurais perceptron multicamada (MLP - *Multi-Layer Perceptron*) e um módulo simbólico que utiliza regras de simples validação léxica das palavras. Os resultados das redes e do módulo simbólico são recebidos na entrada de uma SRN que fará a ponderação destes fatores e terá em sua saída a palavra corrigida. Além disso, há ainda a reorganização da frase com a validação léxica simbólica e de categorias feitas por redes MLP. A correção final é realizada por uma SRN que recebe os resultados das validações e dá como saída a frase corrigida segundo os fatores de entrada.

A frase corrigida é passada para a **parte de casos**, com as palavras mantendo a etiquetagem recebida dos classificadores sintáticos e semânticos básicos e abstratos. Estas informações são utilizadas no parser de segmentação que faz o preenchimento de quadros (*frames*) que compõem os casos. As classes sintáticas e semânticas básicas são usadas para identificar um quadro, enquanto as categorias abstratas são utilizadas para o preenchimento internos dos escaninhos (*slots*) dos quadros. Apesar de todas as características envolvidas, ainda pode haver ambigüidade quanto ao preenchimento de um quadro, mas Wermter e Löchel afirmam que só informações de prosódia permitiriam a solução deste tipo de caso (WERMTER; LÖCHEL, 1996).

A **parte de diálogo** é uma SRN que é treinada para determinar que tipo de ação de diálogo há numa frase. Por fim, o resultado final do sistema é um conjunto de quadros com frases devidamente etiquetadas e classificadas conforme o contexto de diálogo. O sistema todo, portanto, recebe os resultados de um reconhecedor de fala e disponibiliza a organização de uma fala espontânea numa base de conhecimento, disponível para utilização por outros sistemas interativos.

2.2 Regras e Quadros com Análise Estatística

Em meados dos anos 1990, dada a evolução dos sistemas de reconhecimento de fala, antigos sistemas de análise de linguagem natural foram otimizados para acompanhar esta evolução. Um desses sistemas foi o GLR*, que é uma evolução do GLR (*Generalized LR*) que, por sua vez, é o sucessor do LR (*Left-Right*) (LAVIE, 1996a). Este tipo de analisador vem dos anos 1960, onde eram descritas gramáticas simples de avaliação da esquerda para direita, na forma de autômatos determinísticos.

Já o parser GLR é uma extensão do LR que permite a avaliação de gramáticas não-determinísticas. Para tanto, ele usava uma pilha de grafo estruturado (*Graph Strutured Stack* - GSS) que permitia a representação de múltiplas pilhas de parsing, correspondendo a diferentes ações de avaliação de frases.

O parser GLR* foi uma otimização para permitir a análise da fala espontânea, com ênfase na interpretação de palavras desconhecidas e nas limitações de uma avaliação puramente gramatical. Para resolver estes problemas, foi acrescentado um módulo de tratamento de exceção, que processa as palavras retiradas do processo convencional, que antes eram tidas como erradas.

O processamento do GLR* é basicamente a redução das palavras de uma expressão a uma tabela de parser, rejeitando-se os elementos que não se ajustam a estas tabelas. Além disso, há um módulo estatístico para o tratamento de ambigüidades, onde é usado um modelo probabilístico de estados finitos treinado para corrigir a etiquetagem das palavras segundo as ações do parser. Desta forma, o GLR* proporciona uma adequada avaliação da linguagem falada espontânea, conforme comprovado com sua validação no sistema JANUS, que realiza tradução falada interlínguas (LAVIE, 1996a,b).

De forma muito similar ao GLR*, o parser Phoenix também foi construído com uma gramática de regras, que é compilada numa estrutura de grafos chamada Rede de Transição Recursiva (*Recursive Transition Network - RTN*) (MINKER; GAVALDÀ; WAIBEL, 1999; KAISER; JOHNSTON; HEEMAN, 1999). Cada regra cria uma RTN, que, uma vez constituída, tem seus nodos não terminais extraídos e são acrescentados pesos para as conexões dos vértices através de regras heurísticas.

As expressões, antes de serem apresentadas à RTN, têm retiradas as palavras desconhecidas pelo léxico do sistema. Após a avaliação pelas diversas RTNs, são mantidas as redes de maior compatibilidade entre a expressão e seus pesos. As expressões resultantes são então encaixadas em quadros (frames) de uma gramática de casos.

Após, o conjunto de expressões etiquetadas são comparadas com HMMs treinados para reconhecimento de frases, visando a escolha das melhores seqüências semânticas (MINKER, 1998; MINKER; GAVALDÀ; WAIBEL, 1999). Estas expressões avaliadas semanticamente são a saída do parser, as quais podem ser utilizadas posteriormente em sistemas de diálogo.

2.3 Análise de Segmentos com Redes Neurais Artificiais

A análise de segmentos (*chunk parsing*) é baseada na verificação de partes das sentenças faladas (BUØ; WAIBEL, 1999; ZECHNER; WAIBEL, 1998). A análise de segmentos costuma ser apenas um módulo de grandes sistemas de análise da linguagem, mas a seguir será apresentado o FeasPar, que utiliza o processamento de segmentos como base do sistema.

O parser FeasPar faz a geração de quadros (*frames*) de segmentos através de redes

MLP com algoritmo backpropagation. Inicialmente são treinados pela rede conjuntos de quadros chamados *estruturas de características* . Nestas, as sentenças estão devidamente repartidas nos segmentos modelados nos quadros (BUØ; WAIBEL, 1999; BUØ, 1996). Estes quadros construídos manualmente são os padrões de treinamento das redes neurais, cujo objetivo é ensinar ao sistema a construção automática de novos quadros.

O FeasPar possui três grandes módulos, todos constituídos de redes neurais: o segmentador (*chunker*), o etiquetador de características lingüísticas (*Linguistic Feature Labeler* - LFL) e o buscador de relações entre segmentos (*Chunk Relation Finder* - CRF). O segmentador possui três redes neurais que são treinadas para reconhecer números e palavras, frases, segmentos e sentenças. O etiquetador são diversas redes neurais treinadas para classificar características, identificando quais delas pertencem a quais segmentos.

O buscador cria uma rede neural para identificar que tipo de segmento existe em determinada sentença. Uma vez identificados os segmentos, o buscador cria uma outra rede neural para fixar a relação entre eles.

Todas as redes neurais são criadas numa etapa de treinamento, uma vez que a utilização prática se dá posteriormente. Para o nível de segmento, são aplicados o etiquetador e o buscador para, ao final do processo, ser aplicado um algoritmo de busca para preenchimento de um conjunto de possíveis quadros (estruturas de características) para a sentença de entrada. Para decidir qual é o melhor quadro, aplica-se um algoritmo de comparação entre eles.

O algoritmo de busca realiza a montagem dos quadros prováveis a partir de partes dos segmentos, chamados fragmentos. Estes são armazenados durante o processo numa estrutura chamada agenda. A agenda realiza o armazenamento da saída das redes neurais e faz uma composição coerente dos fragmentos, para posterior construção do quadro. Ao final do parsing, a agenda possui o conjunto de quadros que serão avaliados para escolha da melhor estrutura de características da sentença apresentada ao sistema.

2.4 Uso da prosódia em sistemas de compreensão da fala

A prosódia é um dado subliminar à linguagem que expressa a forma com que esta é pronunciada. Por ser uma informação que revela a expressão, diversas situações de análise de ambigüidade podem ser resolvidas com o uso das marcas prosódicas contidas na fala. Este tema tem sido motivo de tema de publicações, como a *Speech Communication* volume 36 (SWERTS; TERKEN, 2002) e conferências, como a *ISCA Speech Prosody 2006* promovida pela International Speech Communication and Association (ISCA).

Desta forma, os sistemas de reconhecimento da fala incorporam, além do reconhecimento dos fonemas ou sílabas, também a identificação das características prosódicas. Assim, são geradas análises da linguagem acrescidas de etiquetas prosódicas (ver seção 4.4). Estas etiquetas auxiliarão na resolução de ambigüidades nos níveis de sintaxe, semântica e pragmática.

2.4.1 Prosódia na sintaxe

No estudo do uso da prosódia na compreensão da linguagem falada, Kompe observou que na análise da sintaxe é possível usar a prosódia basicamente para duas operações: a verificação da acentuação e a otimização da busca da melhor seqüência de palavras (KOMPE, 1997). A acentuação é confirmada pelas etiquetas prosódicas a ela referentes. A seqüência de palavras deve ser pontuada corretamente, estabelecendo as frases sintáticas com vírgulas e pontos em locais apropriados. Para tanto, são utilizadas as etiquetas

de frases prosódicas e pausas identificadas em padrões anteriormente treinados.

Diferentes sistemas podem utilizar as características prosódicas que forem úteis para sua análise sintática. Gallwitz e outros, por exemplo, segmentaram o sinal de fala em quadros de 10ms, testando para cada um 24 características prosódicas (GALLWITZ et al., 2002). Os quadros com suas características são apresentados a uma rede neural para determinação dos limites das frases prosódicas. Estes limites auxiliarão o processo de reconhecimento de palavras por um sistema baseado em HMM, uma vez que este poderá ser modelado a partir das frases prosódicas, e não de um contexto sem a informação sobre a ocorrência de pausas.

O sistema de Gallwitz baseia-se no fato de que os limites sintático-prosódicos ocorrem freqüentemente junto a sons de fundo, silêncio (pausas não-preenchidas) ou pausas preenchidas. Para identificação destes limites foi treinado um sistema HMM para detecção dos eventos sintático-prosódicos, compondo as etiquetas prosódicas com as sintáticas.

A modelagem prosódico-sintática otimiza o reconhecimento das palavras, como demonstrado no sistema de Gallwitz. Uma vez que a construção das frases prosódicas seguem uma regularidade dentro de uma língua, ela pode ser modelada para uma adequada organização das palavras componentes destas frases. Este tipo de processo permite que a identificação das características prosódicas otimizem a análise sintática tradicional.

Numa pesquisa recente, Hasegawa-Johnson e outros otimizaram um reconhecedor de voz incorporando etiquetas prosódicas ao sistema bayesiano já existente (HASEGAWA-JOHNSON et al., 2005). Eles investigaram o acento acústico através do estudo do pitch para confirmar a sílaba tônica das palavras, além de mostrar o relacionamento entre a sintaxe, prosódia e seqüência de palavras.

Para a mostrar a relação entre o acento do pitch e a sílaba tônica transcrita, Hasegawa-Johnson e outros fizeram três sistemas de reconhecimento de acentos. Os modelos usados foram HMM, TDNN (*Time Delay Neural Network* - veja seção 5.4.1.2) e TDRNN (*Time Delay Recursive Neural Network*). O treinamento dos modelos consistiram em verificar regiões de pitch alto (90% das sílabas), baixo (5%) ou questionáveis (5%) para o reconhecimento das sílabas tônicas. O TDRNN mostrou-se o melhor identificador, com uma taxa de erro de 10,2% de um total de 6996 acentos (HASEGAWA-JOHNSON et al., 2005).

O relacionamento da sintaxe com a prosódia foi proporcionado com o desenvolvimento de um modelo bayesiano de bigrama (sistema para previsão da ocorrência de dois termos juntos) para as etiquetas prosódicas. Uma vez criadas as probabilidades de etiquetagem de palavras, foi construído um modelo de linguagem com etiquetas sintáticas e prosódicas. O reconhecimento obtido pelo modelo probabilístico sintático-prosódico não ficou mais que 2% acima do obtido pelo modelo tradicional. Apesar disso, Hasegawa-Johnson e outros apontam um uso promissor deste tipo de abordagem em sistemas de compreensão de fala (HASEGAWA-JOHNSON et al., 2005).

2.4.2 Prosódia na semântica

Segundo Kompe, a prosódia auxilia na análise semântica quanto ao *foco* de uma pronúncia (KOMPE, 1997). O foco ou acento focal refere-se à ênfase dada a uma palavra na sentença conforme sua importância. Outra informação prosódica relevante para a semântica é a inflexão da frase, que pode ter o sentido de uma questão ou afirmação, o que pode mudar completamente o seu sentido (KOMPE, 1997).

No trabalho de Kompe, verifica-se que os dados prosódicos usados na semântica são analisados a nível pragmático. O processamento realizado pela resolução de ambigüidades semânticas mostra-se muito útil para a escolha do contexto de diálogo (KOMPE,

1997).

Trabalhos recentes acerca da investigação das funções da prosódia na pragmática têm analisado o efeito da inflexão da voz na determinação do contexto de significado. Um exemplo é o trabalho de Braga e Marques, o qual demonstra as categorias determinadas pela freqüência da onda fundamental, energia e duração de um segmento de fala. Com base nisso, foram detectadas as modalidades de afirmação, ironia, refutação, emoção e questionamento retórico (BRAGA; MARQUES, 2004).

Outro exemplo da importância da prosódia na pragmática é a análise de itens não lexicais, tais como expressões *uh-huh* e *hmm*, realizado por Niguel Ward. As correlações citadas por Ward vão desde a repetição (*uh-huh-uh-huh*) como indicação de escuta, uma duração maior (*hmmmm*) como sinal de incerteza, maior altura do pitch para indicar o grau de interesse, entre outras características (WARD, 2004).

Outro trabalho que utiliza a prosódia para análise de semântica e pragmática, visando dar robustez a um sistema de diálogo de fala, é o de Zhang e outros. Eles analisam o papel da prosódia para indicar o *foco* e o *contraste* de uma frase. O foco indica uma informação nova de quem fala para quem ouve. O contraste mostra uma contraposição de contextos, do tipo *ocorre assim, **mas** pode ocorrer de outro modo* (ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006).

O sistema de Zhang e outros inicia com um reconhecedor de voz por HMM (ver seção 5.3) e utiliza outro sistema para extração das características prosódicas, tal como pitch, energia e duração dos fonemas. Eles desenvolveram uma técnica chamada *média de acento de pitch*, que faz uma média do pitch detectado em várias sub-bandas de uma análise de multiresolução wavelets. A transformada por wavelets não é citada diretamente no artigo, mas descrevem o uso do filtro Daubechies-4 no processamento (ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006).

Para a análise semântica das palavras utilizadas no sistema, Zhang e outros projetaram uma ontologia para determinar uma hierarquia de categorias semânticas. Assim, foi possível uma comparação do uso das palavras quanto ao contexto em que elas estão inseridas. O cálculo de similaridade leva em conta, além do contexto, a freqüência de uso conjunto de um par de palavras (ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006). O resultado deste processo é a etiquetagem de partes de fala, semelhante à análise de segmentos comentada na seção 2.3.

Por fim, o sistema de Zhang e outros realiza a composição das características prosódicas e da etiquetagem semântica na forma de entrada para uma rede neural TDRNN (já citado na seção 2.4.1) para detectar o foco prosódico das frases. Por outro lado, para teste do contraste da frase foi primeiro aplicado um sistema baseado em conhecimento para selecionar as frases com esta característica e após usado o cálculo de similaridade. Estes sistemas serão compostos num analisador semântico robusto para compreensão de fala, onde o foco auxilia na determinação do contexto e o contraste na estrutura do discurso (ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006).

2.4.3 Prosódia na pragmática - os atos de diálogo

O diálogo tem sido uma área extensamente abordada em sistemas de compreensão da linguagem falada. Portanto, é natural que a análise sintática e semântica seja deixada num segundo plano, uma vez que, neste contexto, a correção ortográfica não é importante. Nesta perspectiva, pode-se utilizar as etiquetas prosódicas diretamente na escolha de um contexto de diálogo, otimizando o processamento através da queima de etapas desnecessárias.

Os contextos da análise pragmática são chamados de atos de diálogo (ver seção 4.4.3). Cada sistema define estes atos, que podem ser, por exemplo: requisição, aceitação, despedida, cumprimento, sugestão, opinião, etc. Os atos podem ser, inclusive, extremamente específicos, como para controles de jogos (HASTIE; POESIO; ISARD, 2002; BOYE; GUSTAFSON; WIRÉN, 2006).

Nöth e seus colegas pesquisadores mostraram as vantagens da unificação das análises de parsing (NÖTH et al., 2002). Eles mostraram que é possível utilizar-se do grafo gerado pelo reconhecedor de palavras com etiquetas prosódicas e processá-lo diretamente em função da sintaxe e da semântica contida nos atos de diálogo.

Neste sistema, os limites de frases são classificados por redes neurais do tipo perceptron multicamada e a busca dos possíveis atos de diálogo é feita no grafo através do algoritmo A*. Para que seja possível encontrar o ato de diálogo certo a equipe de Nöth necessitou implementar um parser probabilístico de n-gramas (ver seção 5.3) para ser um antecipador de conceitos semânticos. Em outras palavras, foi treinado um sistema para identificar uma ocorrência de determinadas expressões que identifiquem contextos. A escolha destas expressões é feita com base na ênfase dada à pronúncia de certas palavras, que vão definir o foco do contexto semântico. Para cada contexto treinado no antecipador, é feita uma gramática de fragmentos que identifica a que atos de diálogo pertencem as frases.

Numa outra abordagem, Stolcke e outros desenvolveram um sistema que identifica atos de diálogo com base numa gramática de discurso, no reconhecimento de palavras e nas características prosódicas da fala (STOLCKE et al., 1998). O reconhecedor de palavras foi treinado com trigramas para reconhecer palavras seguindo a modelagem dos 42 atos de diálogo.

As características prosódicas utilizadas são a duração, pausa, onda fundamental, energia e gênero (masculino/feminino). Estas características foram classificadas em árvores de decisão, cujos nodos possuem estatísticas de ocorrências de determinada característica prosódica.

Essas três abordagens têm suas probabilidades comparadas ao final do processamento. A maior probabilidade define o ato de diálogo que será associado à pronúncia de entrada do sistema.

Num sistema baseado inteiramente em redes neurais TDNN (veja seção 5.4.1.2), Kipp mostrou que é possível a inferência de atos de diálogo a partir de etiquetas prosódicas (KIPP, 1998). Neste sistema, há um conjunto de redes TDNN, cada uma destinada a representar determinado ato de diálogo.

No sistema de Kipp, a entrada para as redes seria uma codificação de uma janela de etiquetas prosódicas de dada sentença. Como as redes possuem um número fixo de entradas, é necessário um esquema de janelamento, onde são avaliadas apenas as etiquetas constantes na janela. Estas etiquetas são codificadas e treinadas para o reconhecimento de determinado ato de diálogo.

O reconhecimento final do ato de diálogo é dado por uma rede especialmente treinada para identificar a maior probabilidade dentre as redes treinadas. Esta rede dará a saída final para a identificação do ato de diálogo.

2.5 Sistemas de Diálogo de Fala

Sistemas de compreensão de fala são uma primeira parte de sistemas de diálogo de fala, estes possuidores de uma segunda parte responsável pela elaboração de respostas e

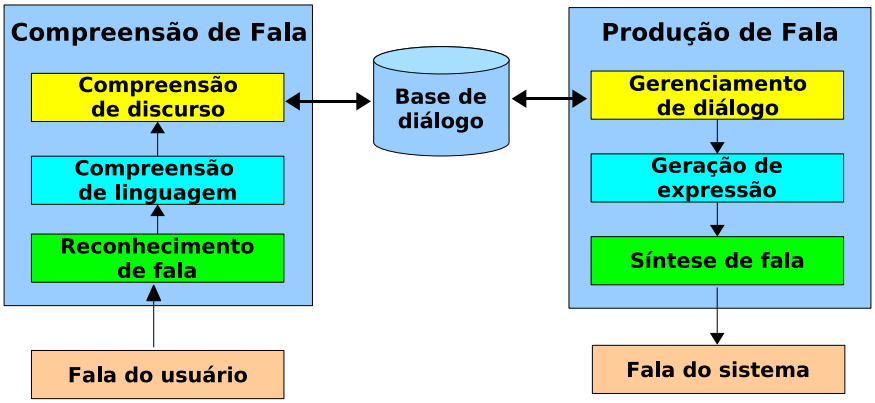

Figura 2.2: Sistema de diálogo falado.

síntese de voz. Segundo (HIGASHINAKA; SUDOH; NAKANO, 2006), um sistema de diálogo falado é composto de dois grandes módulos, a compreensão de fala e a produção de fala, conforme pode ser observado na figura 2.2.

Um exemplo recente que incorpora características tradicionais de sistemas de diálogo de fala é o de Higashinaka e outros. Eles argumentam que um sistema de diálogo deve reconhecer a intenção da fala através do contexto, ou seja, do estudo do discurso. Neste sistema, eles utilizam uma estrutura de escaninhos (*slots*) para organização de características da fala e da linguagem, como utilizado no sistema SCREEN (ver seção 2.1). Estas são características do modelo acústico, de linguagem e resultante da análise do discurso (HIGASHINAKA; SUDOH; NAKANO, 2006).

Neste sistema, Higashinaka e outros acoplaram sistemas de reconhecimento e síntese de outros autores. Elaboraram um modelo de trigramas para uma base extraída de um serviço telefônico de informações meteorológicas japonês. Após escolhidas as falas, estas foram transcritas e etiquetadas por atos de diálogo (como na seção 2.4.3) e conceitos similares. Então, as frases foram treinadas em um modelo bayesiano conhecido como Transdutor de Estados Finitos, que visa a construção de uma gramática estatística, semelhante à RTN comentada na seção 2.2. O resultado da aplicação da gramática estatística é a catalogação de atos de diálogo encaixados em escaninhos de conceitos, segundo sua estrutura de organização (HIGASHINAKA; SUDOH; NAKANO, 2006).

A resposta do sistema é organizada por *regras de compreensão de discurso*, onde cada ato de diálogo preenchido em determinado escaninho possui uma resposta-padrão previamente preparada. A resposta é sintetizada de volta ao usuário. Caso falte algum dado, como o nome de uma cidade, o sistema questiona procurando completar o escaninho. Caso o usuário faça pausas, o sistema emite sons não lexicais como *uh-huh* (HIGASHINAKA; SUDOH; NAKANO, 2006).

2.6 Um modelo biologicamente plausível?

Neste capítulo, pudemos perceber que não são comuns os sistemas que atendem desde o reconhecimento da fala até o processamento de diálogos. Apenas recentemente vemos sistemas completos de diálogo falado, com reconhecimento e síntese de fala, como o de Higashinaka citado na seção anterior.

Nesses sistemas mais completos, a prosódia aparece apenas como um complemento no processamento do reconhecimento de voz e identificação de atos de diálogo. No re-

conhecimento, a prosódia permite a identificação das sílabas tônicas das palavras, auxiliando na identificação do léxico. Na definição de contextos, realiza a percepção de uma palavra que atua como foco da frase, contextualizando-a. Apesar destas aplicações da prosódia, nenhum dos trabalhos estudados menciona um modelo biológico que embase seu desenvolvimento.

Dentre os modelos citados neste capítulo, nenhum deles está indicado como sendo biologicamente plausível, ou seja, inspirado no comportamento neurológico do ser humano para a compreensão da linguagem falada. Neste sentido, o capítulo seguinte busca uma motivação biológica para o desenvolvimento de um novo modelo computacional para a compreensão da fala.

3 MODELO NEUROCOGNITIVO

Um sistema computacional em geral realiza a representação de uma modelagem do mundo real. No caso da compreensão da fala, propõe-se como base um modelo neurocognitivo de frases escutadas. Especificamente, o modelo aqui utilizado é o de Angela Friederici (FRIEDERICI, 1995, 2002), que descreve a temporalidade da compreensão de frases a partir de sinais e imagens do cérebro.

O objetivo do modelo de Friederici foi mapear no cérebro o processamento e compreensão sintática e semântica de frases escutadas. Para tanto, dada a audição de uma frase, localizam-se as áreas do cérebro ativadas no decorrer do tempo. Com base nesta localização, é possível identificar as áreas responsáveis pelo processamento de cada etapa da compreensão das frases escutadas.

Segundo Friederici, há duas grandes correntes de modelos psicolingüísticos de compreensão da informação fonológica. Uma sustenta que ocorre um processamento seqüencial, sendo primeiro a análise sintática e depois a semântica. Outra argumenta no sentido que ocorre uma interação entre as análises (FRIEDERICI, 2002).

O modelo seqüencial defende que o processamento da linguagem inicia com a estruturação sintática, a partir das categorias das palavras, para posteriormente ocorrer a análise semântica. Já a visão interacionista sustenta que acontece um relacionamento entre sintaxe e semântica desde o processamento da audição, o qual prossegue durante todo o processo de compreensão da linguagem (FRIEDERICI, 2002).

O modelo de Friederici não segue exatamente nenhuma das correntes psicolingüísticas, uma vez que constata a ocorrência dos dois modelos, em diferentes momentos, nas reações cerebrais decorrentes do processamento da audição da linguagem. Em outras palavras, o modelo neurocognitivo baseia-se em exames laboratoriais do cérebro realizados durante o processamento da audição de frases.

A relação da linguagem com exames no cérebro iniciou com Paul Broca (o sobrenome pronuncia-se *brocá*), que mostrou, em 1861, que a capacidade de produção da fala era localizada no hemisfério esquerdo, mais exatamente no giro frontal inferior (SCOTT; WISE, 2003; FOZ et al., 2005). Posteriormente, Karl Wernicke identificou em 1874 que no primeiro giro temporal posterior, também no hemisfério esquerdo, como sendo uma área relacionada com a percepção da fala, ao qual associou estruturas cerebrais que relacionariam esta área àquela descoberta por Broca (SCOTT; WISE, 2003; FOZ et al., 2005). A figura 3.1 indica as áreas de Broca e Wernicke no hemisfério esquerdo do cérebro.

No século 20 cabe citar a contribuição de Luria, entre os anos 1940 e 1980, com a descrição de diversos tipos de distúrbios da fala e da linguagem (afasia), e de Geschwind, que apontou, em 1965, uma área do giro angular do cérebro como sendo responsável pela compreensão da linguagem escrita (FOZ et al., 2005). Nos anos 1990 e início do século

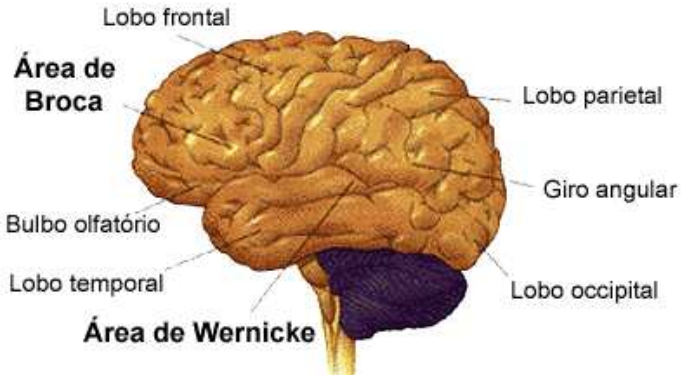

Figura 3.1: Áreas de Broca e Wernicke (LETHBRIDGE, 2001).

21 os estudos neurocognitivos obtiveram um avanço significativo permitido pelas novas técnicas de análise disponíveis.

Segundo Scott e Wise, com o advento da neuropsicologia cognitiva nos anos 1960, foram criados modelos lingüísticos com base em disfunções cerebrais. As disfunções cerebrais tinham mais relevância do que a tentativa de associar uma área cerebral a uma função regular. Nos anos 1970 foram possíveis avanços através das análises de imagens de raios-x, apesar de sua baixa resolução. Uma melhor qualidade de imagens foi obtida posteriormente com a tomografia computadorizada, onde a pessoa em análise bebe um líquido radioativo conhecido como *contraste*, e a radiação emitida é captada por uma câmara ligada a um computador. A modalidade mais utilizada atualmente deste tipo de exame é a tomografia por emissão de pósitrons (PET - *Positron Imaging Tomography*).

Segundo Oliveira Filho, o PET consiste na aplicação de um isótopo emissor de pósitrons na pessoa que será submetida ao exame (OLIVEIRA FILHO, 2005). Dependendo da molécula do corpo que se deseja analisar, utiliza-se um líquido de contraste baseado em glicose etiquetada com o isótopo que reage melhor com as moléculas do tecido de interesse (SCOTT; WISE, 2003; OLIVEIRA FILHO, 2005). Como a glicose é fonte primária de energia das células, ela é absorvida por elas e os pósitrons colidem com os elétrons livres encontrados. Dessa colisão resultam fótons (raios gama) que são captados pela câmara na qual a pessoa fica inserida. A organização e intensidade dos fótons emitidos é calculada por computador, o qual realiza a construção da imagem (OLIVEIRA FILHO, 2005).

Como o PET identifica as células, ele pode ser utilizado para realizar uma varredura do fluxo de sangue nos órgãos. Scott e Wise explicam que o PET é utilizado para identificar a atividade dos neurônios no cérebro. Uma vez que estes aplicam muita energia na ativação sináptica, o fluxo de sangue utilizado para suprir as necessidades dos neurônios indica a área que está sendo utilizada em dado momento (SCOTT; WISE, 2003).

Outra técnica de neuroimagem largamente utilizada é a ressonância magnética (MRI - *Magnetic Resonance Imaging*), que aperfeiçoou muito a qualidade das imagens, inclusive pela capacidade de registrar fatias de imagens em qualquer orientação (SCOTT; WISE, 2003). Segundo Amaro Jr. e Yamashita, a ressonância magnética é uma técnica de neuroimagem que realiza três etapas: alinhamento dos átomos, excitação do hidrogênio e detecção de radiofreqüência. Através de um campo magnético intenso é realizada a orientação dos átomos. Após, gera-se uma onda eletromagnética de mesma freqüência do átomo de hidrogênio, ou seja, 63,8 MHz. Por fim, as imagens são geradas pela resposta

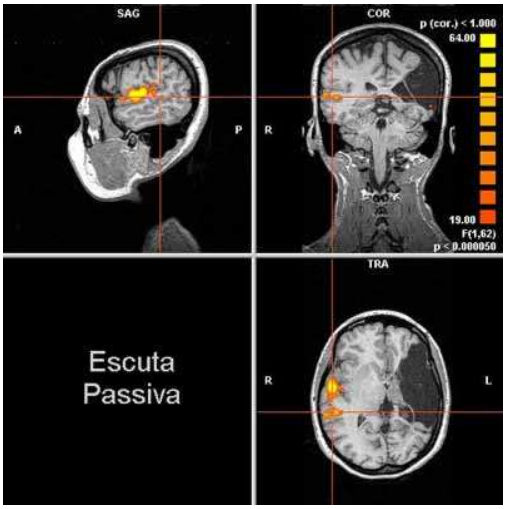

Figura 3.2: Exemplo de fMRI para audição (ROSSET et al., 2005).

de freqüência emitida pelos átomos de hidrogênio. Dependendo da intensidade desta resposta, geram-se tons de cinza que formam a imagem (AMARO JUNIOR; YAMASHITA, 2001; COVOLAN et al., 2004).

No sentido de realizar análises funcionais, a ressonância magnética pode ser associada a determinada atividade, como a audição de uma frase, por exemplo. Assim, a ressonância magnética funcional (fMRI - functional MRI) responde não ao hidrogênio, mas à célula do sangue (hemoglobina). A hemoglobina saturada de oxigênio possui uma resposta diferente da não-saturada, o que permite identificar que áreas do cérebro estão consumindo mais oxigênio, ou seja, possuem maior atividade (COVOLAN et al., 2004). Como a ressonância não envolve radiação, ela pode ser aplicada diversas vezes a uma mesma pessoa, inclusive desenvolvendo as atividades na qual se deseja analisar (AMARO JUNIOR; YAMASHITA, 2001; SCOTT; WISE, 2003). Um exemplo de fMRI de audição obtida por pesquisadores da Universidade de São Paulo é apresentada na figura 3.2.

Talvez a técnica de análise do comportamento cerebral mais disseminada seja o eletroencefalograma (EEG). Ele foi criado pelo psiquiatra alemão Hans Berger, tendo o primeiro registro efetuado em 1929 (ARAÚJO; CARNEIRO; BAFFA, 2004). Segundo Aniela França, os sinais elétricos detectados por eletrodos colocados no couro cabeludo permitem a detecção do comportamento do cérebro enquanto a pessoa em análise realiza alguma atividade de interesse no estudo. Os sinais captados são pré-processados pela *promediação*, que é a soma dos sinais captados. Isso visa a eliminação de ruídos vindos de outras fontes de energia ou estímulos fora da área de interesse na análise. A onda resultante da promediação é chamada de potencial relativo a evento (ERP - *Event-Related brain Potential*). Há ondas ERPs relacionadas a fenômenos lingüísticos, tais como: negatividade anterior esquerda precoce (ELAN - *Early Left Anterior Negativity*), que detecta eventos entre 125 e 180 ms, onde pode indicar um erro de interpretação de classe de palavras; a negatividade anterior esquerda (LAN - *Left Anterior Negativity*), que ocorre entre 300 e 500 ms e indica erros morfosintáticos; negatividade 400 (N400), que tem sua amplitude aumentada quando ocorre um erro semântico na compreensão do sentido de uma frase; positividade centroparietal tardia 600 (P600), que indica aos 600 ms o entendimento de um erro gramatical; componente positivo tardio (LPC - *Late Positive Component*), que ocorre entre 500 e 800 ms e acusa a percepção do erro na formação morfológica de pa-

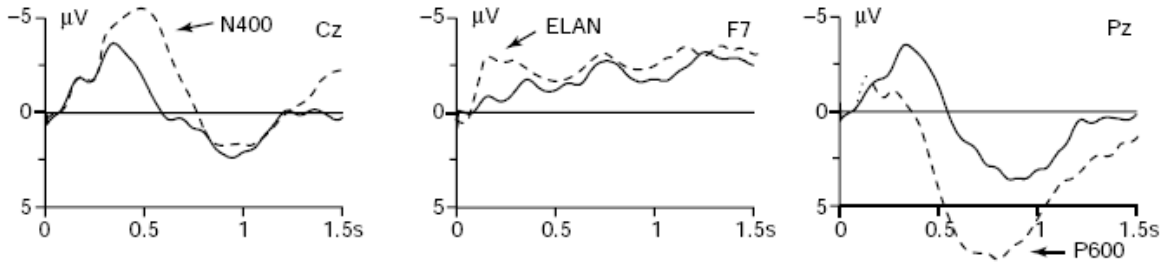

Figura 3.3: Sinais N400, ELAN e P600 (FRIEDERICI, 2002).

lavras (FRANÇA, 2005). Exemplos de sinais N400, ELAN e P600 são apresentados na figura 3.3.

Por fim, uma técnica de exame que também é utilizada para verificação neurocognitiva é a magnetoencefalografia (MEG). Enquanto o EEG mede os sinais elétricos, o MEG detecta o campo magnético gerado por estes sinais ao passarem pelos neurônios. O MEG foi criado no final dos anos 1960 por David Cohen e em 1972 foi usado para registrar sinais alfa emitidos pelo cérebro (ARAÚJO; CARNEIRO; BAFFA, 2004). Essa técnica possui excelente precisão temporal e espacial (em camadas superficiais), mas tem como desvantagens o fato de alguns neurônios produzirem campo magnético nulo e do material, cujos sensores supercondutores necessitam de refrigeração especial (ARAÚJO; CARNEIRO; BAFFA, 2004). Os sinais ERP também são obtidos pelo MEG, mas somente até os ELAN, pois os demais não retornam uma localização espacial satisfatória (FRIEDERICI; KOTZ, 2003).

3.1 Modelo Neurocognitivo de Processamento da Audição de Frases

Com base em exames PET, fMRI, EEG e MEG, Friederici propôs um *Modelo Neurocognitivo de Processamento da Audição de Frases (MNPAF)* (FRIEDERICI, 2002). Este modelo foi construído com base em análises dos campos cerebrais que reagem segundo determinadas condições da linguagem apresentada a voluntários - em geral jovens - que se submetem aos exames.

Os exames são procedidos colocando-se o voluntário no equipamento (ou ligado a ele, no caso do ERP) e ditando-se frases a ele. As frases são divididas em corretas, sintaticamente incorretas, semanticamente incorretas, ou ambas incorretas (FRIEDERICI; KOTZ, 2003). Para cada tipo de frase, verificam-se quais partes do cérebro são ativadas no decorrer do tempo.

Por frases incorretas sintaticamente entenda-se a retirada de um elemento que permite a perfeita compreensão, por exemplo, a frase *O sorvete estava no * comido*. Uma frase incorreta semanticamente seria *O vulcão foi comido*. Já uma frase com as duas formas de erro (sintático e semântico) seria *A porta trancada estava no * comido*. E uma frase correta seria *O sorvete foi comido*. As frases marcadas com o símbolo * indicam a retirada de um substantivo naquele ponto da frase, propositalmente antes do verbo no particípio. Todas as frases dos testes de Friederici foram construídas na voz passiva para que o verbo mais importante ficasse ao final da ordem de audição. As frases são apresentadas alternadamente para evitar a previsão da construção do próximo exemplo. Estes exemplos estão relatados em (HAHNE; FRIEDERICI, 2002).

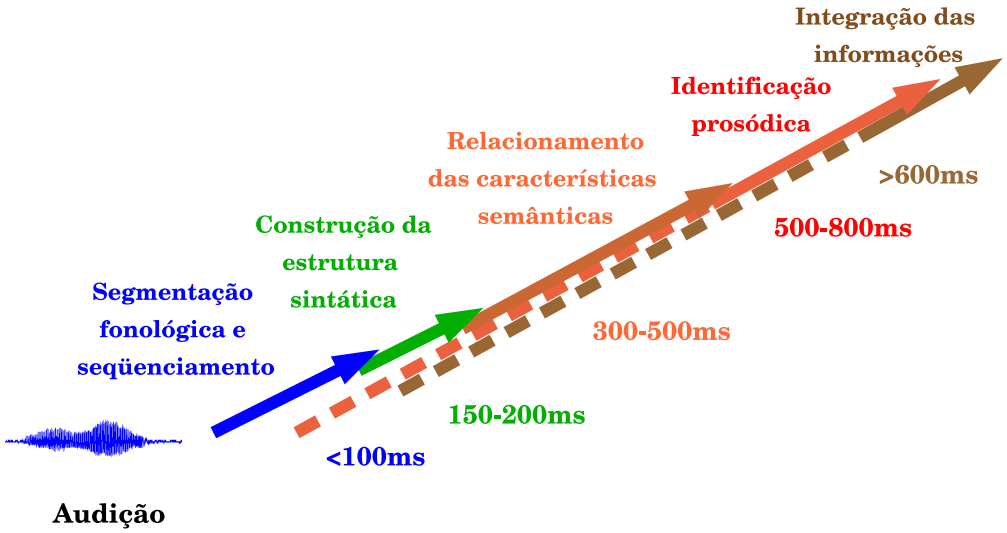

Figura 3.4: Seqüência ampliada do modelo neurocognitivo da audição de frases.

Além do registro gramatical, a ativação da prosódia no cérebro também é avaliada através de exames ERPs e fMRIs. Na pesquisa procedida por (HERRMANN et al., 2003) foram utilizados como base testes de verificações sintáticas semelhantes aos citados anteriormente, mas com a fala monótona (também conhecido como *achatamento* do pitch). Este efeito é obtido extraindo-se a onda fundamental (F0), colocando-a numa freqüência constante, por exemplo a 180 Hz, e resintetizando a fala (HERRMANN et al., 2003). Uma investigação sobre a reação a uma fala positiva (feliz), negativa (triste) e neutra - sem erros sintáticos - foi conduzida por Sonja Kotz e outros, usando exames fMRIs (KOTZ et al., 2003). Outra pesquisa, realizada por Martin Meyer e outros, também com exames fMRIs, verifica a ativação das áreas cerebrais para processamento e produção de informação prosódica (MEYER et al., 2004). Neste tipo de exame, são apresentadas frases corretas em fala normal, monótona e abafada. Este último tipo de fala é sintetizado com som abafado, como se fosse escutado atrás de uma porta. Este efeito foi obtido alterando-se constituintes prosódicos como duração, amplitude e entonação (MEYER et al., 2004).

Através de suas experiências, Friederici mapeou quais partes do cérebro eram ativadas no tempo, dados os exames aplicados. No seu modelo neurocognitivo Friederici dividiu o processamento das frases ouvidas em 4 grandes fases onde são relacionadas funções com áreas do cérebro correlacionadas com seu tempo de ativação (veja figura 3.4): segmentação fonológica e seqüenciamento (até 100 ms); construção da estrutura sintática (entre 100 e 300 ms); definição das relações semânticas e papel temático (entre 300 e 500 ms); integração sintática-semântica-prosódica (que compõe-se a partir da análise sintática e ocorre entre 500 e 1000 ms) (FRIEDERICI, 2002; FRIEDERICI; KOTZ, 2003; FRIEDERICI; ALTER, 2004).

Como se pode perceber, no modelo de Friederici a análise da sintaxe precede a semântica. Apesar disso, os resultados da sintaxe não atuam sobre a semântica, mas os resultados de ambas as análises são comparadas ao final do processo, na fase de integração (FRIEDERICI, 2002).

Recentes pesquisas indicam que a este modelo neurocognitivo deve ser acrescentada a descrição do processamento da prosódia. Esta seria mais uma fase que começa a ser composta aos 40 ms, ocorre aos 500 ms e integra-se com as demais categorias aos 800 ms

(ECKSTEIN; FRIEDERICI, 2005).

O fluxo de processamento do modelo neurocognitivo inicia, então, com a análise acústica até 100 ms (fase 0), passando pela análise sintática entre 100 e 300 ms (fase 1), após pela análise semântica entre 300 e 500 ms (fase 2), agora mais a análise prosódica aos 500 ms e por fim a integração de todas as etapas entre 500 e 1000 ms (fase 3). Nas próximas seções serão analisadas cada uma das etapas descritas.

3.1.1 Fase 0 - análise acústica

A primeira fase, chamada por Friederici de *fase 0*, foi identificada em exames ERP por sinais N100 (ondas negativas aos 100 ms) (FRIEDERICI, 2002). Nesta fase ocorre a análise acústica inicial e a segmentação do sinal. Em estudos específicos sobre esta fase, verificou-se que a resposta inicial do córtex auditivo primário ocorre entre 15 a 19 ms após a audição, é distribuído para as partes laterais da área de Heschl (giro temporal transverso anterior) entre 25 e 40 ms e a detecção do pitch ocorre entre 60 e 80 ms (GUTSCHALK et al., 2004). Segundo Gutschalk e outros, ainda não há certeza de como ocorre a formação do pitch, que tanto pode ser a representação da integração de canais de freqüência de uma informação obtida por intervalos de tempo, como o cálculo de um valor de pitch específico e sua intensidade (GUTSCHALK et al., 2004).

Sobre o papel do pitch no processamento da compreensão da fala, Friederici e outros de sua equipe do Instituto Max Planck argumentam que, quando é usado como informação afetiva, seu processamento é feito apenas no hemisfério direito mas, quando envolvido em tarefas lingüísticas, é também processado pelo hemisfério esquerdo (HERRMANN et al., 2003; FRIEDERICI; ALTER, 2004). Segundo Friederici e Alter, a importância da análise do pitch reside na necessidade de identificar as informações de entonação da fala, ou seja, suas características prosódicas para a compreensão da fala (FRIEDERICI; ALTER, 2004).

A análise da prosódia é essencial para o processamento da linguagem falada, uma vez que permite a identificação da acentuação e entonação de uma frase (FRIEDERICI; ALTER, 2004). A intensidade e ritmo do pitch possibilita a detecção de segmentos da fala, o que permite, por exemplo, a identificação de pausas e palavras (análise léxica). O reconhecimento de palavras auxiliado pelo pitch é verificado pelo sinal P200 nos exames ERPs, que ocorre entre 200 e 280 ms (FRIEDRICH et al., 2004). Constata-se ainda que o próprio pitch é um elemento à parte no processamento da audição de frases, uma vez que se verifica também a ativação do sinal P350 quando há erro de entonação (FRIEDRICH et al., 2004).

Friederici e Alter sustentam que há participação da prosódia não somente na segmentação, mas também na correção sintática e na análise semântica (FRIEDERICI; ALTER, 2004). A prosódia auxilia na resolução de ambigüidades através da informação de entonação das palavras e na análise semântica através do apoio na decisão da hierarquia papéis temáticos (definição de contextos) (FRIEDERICI; ALTER, 2004). A descrição da influência prosódica na análise da linguagem é aprofundada na seção 3.1.5.

3.1.2 Fase 1 - construção da estrutura sintática

A verificação sintática das frases escutadas é detectada por exames PET, fMRI e ERP. Neste último, a percepção de erros sintáticos é verificada no sinal ELAN, que ocorre entre 150 e 250 ms, e sua tentativa de correção é indicada no sinal P600 (FRIEDERICI; KOTZ, 2003; FRISCH; HAHNE; FRIEDERICI, 2004). O ELAN indica erro de estrutura frasal e a correção do erro se dá aos 600 ms porque há a espera do resultado da análise semântica

(ver seção 3.1.3), a qual auxiliará na solução.

Através de exames fMRI constatou-se que o processamento sintático é concentrado na área de Broca (HEIM; OPITZ; FRIEDERICI, 2003). Friederici, estudando esta área, observou que diferentes pontos são ativados conforme a complexidade da análise sintática (FRIEDERICI, 2004). A capacidade de estruturar hierarquicamente a sintaxe e realizar relações entre as hierarquias é o que distingue a linguagem humana da capacidade de comunicação de outros primatas (FRIEDERICI, 2004). Os diferentes pontos de ativação da área de Broca parecem refletir a capacidade de realizar distantes relacionamentos entre hierarquias sintáticas (FRIEDERICI, 2004).

Sandra Vos e Angela Friederici constataram que pessoas que possuem rapidez na leitura possuem maior facilidade de reorganização da estrutura sintática de frases incorretas (VOS; FRIEDERICI, 2003). As pessoas com lentidão de leitura não ativaram o sinal P600, que indica a reestruturação da sintaxe. Os leitores rápidos, por outro lado, ativaram o sinal P345 (em torno dos 350 ms) e algumas vezes o P600 durante a resolução dos conflitos sintáticos.

Vos e Friederici levantam a hipótese de que os leitores rápidos constroem uma estrutura para resolução de problemas, enquanto leitores lentos são dependentes de contextos conhecidos para solucionar as contradições sintáticas (VOS; FRIEDERICI, 2003). Estes dados são confirmados por outros testes, onde os erros de estrutura de frase são detectados na fase 1 (sinal ELAN) e implicam na necessidade de correção, dispensando a análise semântica (HAHNE; FRIEDERICI, 2002; FRISCH; HAHNE; FRIEDERICI, 2004; ROSSI et al., 2005).

Os dados de Vos e Friederici são também complementados por um estudo comparativo entre diferentes leitores procedido por Ina Bornkessel e outros (BORNKESSEL; FIEBACH; FRIEDERICI, 2004) para processar estruturas ambíguas e complexas. Foi constatado que os leitores lentos tiveram maior tempo gasto para realizar o processamento. Isso foi verificado pelo sinal N400, que indica o uso da análise semântica para resolução, o que não ocorre nos leitores rápidos (BORNKESSEL; FIEBACH; FRIEDERICI, 2004). Isso embasa a hipótese de Bornkessel e outros de que os leitores rápidos constroem uma estrutura específica para solucionar problemas sintáticos.

Nos exames relativos ao processamento temporal sintático, identificou-se que a categoria da palavra é verificada antes do gênero, o que é compatível com o modelo psicolingüístico seqüencial (HEIM; OPITZ; FRIEDERICI, 2003). Neste sentido, Sonja Rossi e outros observaram que a verificação de categoria sintática é um processo que ocorre antes da análise de concordância sujeito-verbo (ROSSI et al., 2005). Isso acontece porque, caso a categoria não se encaixe na organização da frase, considera-se como um erro de estrutura (sinal ELAN seguido do P600) e portanto não chega a ser analisada a concordância (ROSSI et al., 2005).

Os exames cerebrais, principalmente ERPs, mostraram, portanto, que há um processo natural e independente de avaliação sintática da audição de frases que ocorre na área de Broca. Esta avaliação aparentemente se dá de forma hierárquica, onde o topo do processo inicia com a compatibilização das categorias sintáticas na estruturação de uma frase.

Um erro de estrutura é tão grave que pode inviabilizar a análise semântica, como no exemplo visto anteriormente: *O sorvete estava no comido*. Neste caso, a falta de um elemento sintático inviabiliza até mesmo a avaliação do contexto. Ou seja, há casos em que a questão da estrutura deve ser resolvida antes do contexto. Por outro lado, situações problemáticas que não envolvam estrutura, como casos de ambigüidade, vão ser auxiliadas pelo contexto. As definições de contexto são avaliadas pela análise semântica, como será

visto na seção a seguir.

3.1.3 Fase 2 - definição das relações semânticas

A percepção de erros semânticos pode ser detectada pelo sinal ERP N400 , o qual ocorre em torno de 400 ms após a audição de palavras que não podem ser integradas ao contexto da frase (FRIEDERICI, 2002). Também são identificados conflitos de gênero, segundo (HEIM; OPITZ; FRIEDERICI, 2003), e erros de concordância nominal e verbal, através do sinal ERP LAN, que abrange entre 450 e 650 ms, segundo (ROSSI et al., 2005). Além dos exames EEG (sinais ERP citados), alguns testes também envolvem PET e fMRI para auxiliar na localização espacial da origem do processamento semântico.

Sabe-se que a análise semântica é lateralizada no hemisfério esquerdo do cérebro, ocorrendo nas redes neurais no meio e parte posterior do giro temporal superior, no giro temporal médio e ainda na área de Broca (FRIEDERICI; ALTER, 2004). Como a análise sintática também ocorre na área de Broca, observa-se que as avaliações sintática e semântica concentram-se, portanto, no hemisfério esquerdo do cérebro.

Nesta fase 2 são procurados os contextos das frases que obtiveram sucesso na avaliação da estrutura sintática. Uma vez bem estruturadas, as frases são avaliadas quanto ao gênero, à concordância e ao contexto semântico das palavras envolvidas.

Um erro tipicamente analisado nesta fase é o de gênero, como no exemplo *Ela dirige o terra*, cujo erro ativa tanto o sinal N400 como o LAN e posteriormente o P600, indicando a tentativa de correção (FRISCH; HAHNE; FRIEDERICI, 2004). Estes sinais também são ativados no caso de erros de concordância, como o exemplo de caso verbal *Ele sabia que o químico e o físico emigrou*, provocam um conflito na determinação dos contextos temáticos que definiriam a compreensão perfeita da frase (FRISCH; HAHNE; FRIEDERICI, 2004; ROSSI et al., 2005).

As frases que são submetidas à análise semântica são corretamente estruturadas, caso contrário não seriam sequer avaliadas quanto ao contexto, como foi visto na seção 3.1.2. Um exemplo já citado de erro semântico é *O vulcão foi comido*, onde a estrutura está correta, mas o relacionamento proposto pela frase não indica um contexto válido. Este tipo de erro ativa o sinal N400, que é reconhecido por indicar uma falha no reconhecimento semântico, e o P600, que aponta uma tentativa de correção do erro detectado (FRIEDERICI, 2002; FRISCH; HAHNE; FRIEDERICI, 2004; ROSSI et al., 2005).

A análise semântica se dá, segundo os estudos neurocognitivos observados, nos casos de audição de frases corretamente construídas. Uma vez identificadas as categorias sintáticas, são analisadas as relações entre as palavras e os contextos pertinentes às categorias das palavras e suas relações.

3.1.4 Fase 3 - integração sintática-semântica-prosódica

A fase 3 realiza a integração de todas as análises. A necessidade de correção de qualquer fase é efetivada nesta, através do sinal P600, que se dá por volta dos 600 ms após a audição. Esta fase aparentemente faz uma reanálise estrutural provocada por vários fatores como problemas de estrutura sintática, concordância ou definição de contexto semântico (FRIEDERICI, 1995). Friederici sugere que esta fase, ao contrário das primeiras, é menos dependente do tempo e pode ocorrer em paralelo, recebendo informações tanto sintáticas como semânticas (FRIEDERICI, 1995).

Stefan Frisch e outros sugerem que a correção sintática realizada nesta fase pertence a dois tipos: estrutura da frase e argumento da estrutura (FRISCH; HAHNE; FRIEDERICI, 2004). A correção de estrutura é derivada de problemas de organização ainda na fase 1,

proveniente da ordem de entrada dos termos lexicais (FRISCH; HAHNE; FRIEDERICI, 2004). Já a correção do argumento da estrutura ocorre em função da incompatibilidade de termos mal encaixados dentro de uma estrutura esperada (FRISCH; HAHNE; FRIEDERICI, 2004).

Como nesta fase não há maior detalhamento em relação aos exames neurocognitivos, em geral são levantadas hipóteses de como ocorre o processamento cerebral, com base em modelos psicolingüísticos. A descrição que freqüentemente ocorre é a interação lingüística como responsável pelo desenvolvimento de estruturas específicas para o tratamento da linguagem. Esta interação ocorre tanto por causa da leitura e escrita quanto a outras funções motoras (STOWE et al., 2004).

Uma vez constituída a estrutura de avaliação da linguagem através da interação, esta serve como parâmetro para correção. Enquanto não há uma estrutura consolidada para a base sintática, aparentemente o resultado da análise semântica procura auxiliar na solução, como visto na seção 3.1.2. Mas uma vez constituída a estrutura sintática, ela permite uma correção mais eficiente em termos de gênero e concordância.

Por outro lado, também não há dados para avaliação de como seria o processamento diante de uma linguagem organizada hierarquicamente. Como comentado na seção 3.1.2, já existem análises sobre a complexidade de relacionamentos hierárquicos da linguagem humana. Mas até o momento não há análises sobre como se constituem estes relacionamentos, mas são buscados outros elementos que permitam uma melhor interpretação deste quadro, tal como o estudo da prosódia sobre a sintaxe e a semântica.

3.1.5 Identificação prosódica - uma nova fase?

Por estar sendo tratada aqui a linguagem falada, o processamento do sinal prosódico torna-se um importante diferencial. Observa-se que a prosódia é o fio que conduz uma coerente conexão entre todas as fases de processamento da linguagem, que vai da análise léxica à semântica. Na investigação de como o cérebro realiza a segmentação e identificação léxica, Claudia Friedrich e outros observaram que o pitch é codificado de forma independente da segmentação da informação, mas é usado para indicar as representações léxicas (FRIEDRICH et al., 2004).

Segundo Herrmann e outros, testes realizados variando-se a prosódia, de normal a monótona, varia também o sinal ELAN, indicando falhas na análise sintática e, conseqüentemente, mostrando uma dependência entre prosódia e sintaxe (HERRMANN et al., 2003). Eles constataram que, pelo fato de apresentarem uma fala monótona, o processamento exige uma maior atividade do hemisfério direito do cérebro para compensar a falta da prosódia normal e permitir a resolução da linguagem, que ocorre no hemisfério esquerdo (HERRMANN et al., 2003).

Confirmando os dados de Herrmann, mas num estudo usando fMRI, Martin Meyer e outros procederam um exame onde testaram, além da fala normal e monótona, uma fala abafada (MEYER et al., 2004). Eles verificaram que, mesmo com a codificação prosódica realizando-se no hemisfério direito, o processamento do contorno da entonação de frases faladas é feito no hemisfério esquerdo. Foi verificado que o lado direito do cérebro processa uma prosódia não-lingüística, enquanto o lado esquerdo codifica a prosódia que está envolvida em frases. Além disso, a área que a prosódia envolve no hemisfério esquerdo abrange não somente a área de processamento da linguagem, mas também outras que envolvem articulações sensório-motoras e áudio-motoras, sugerindo que existem relacionamentos da compreensão da linguagem com outros sentidos, além da audição (MEYER et al., 2004; STOWE et al., 2004).

Angela Friederici e Kai Alter também constataram que a análise prosódica inicia no hemisfério direito e é concluída no esquerdo (FRIEDERICI; ALTER, 2004). O lado direito teria a função de segmentar e codificar o sinal prosódico que é utilizado pelo lado esquerdo, onde se realiza a interpretação lingüística (FRIEDERICI; ALTER, 2004).

Os estudos de Herrmann, Meyer e de Friederici e Alter são confirmados por Marc Pell que, analisando a prosódia em pacientes com danos no hemisfério direito e outros no esquerdo, constatou resultados semelhantes (PELL, 2006). Os pacientes que trabalhavam melhor com o lado esquerdo do cérebro conseguiam compreender melhor a expressão verbal, enquanto os que tinham o lado direito funcional entendiam mais a fala emocional que a verbal (PELL, 2006).

Em testes relacionando a prosódia com a sintaxe, Korinna Eckstein e Angela Friederici identificaram um sinal de negatividade anterior direita (RAN - Right Anterior Negativity) aos 500 ms que ocorreu dada uma alteração da entonação. Como isso se manifestava tanto em frases sintaticamente corretas como erradas, a RAN é proposta como uma identificação do processamento prosódico (ECKSTEIN; FRIEDERICI, 2005). A integração com os demais componentes é sinalizada por uma positividade aos 800 ms (P800). Todo o processamento tem indicativos de ser puramente prosódico, ou seja, não interfere ou há interferência da análise sintática.

Annett Schirmer e outros constataram que, quanto mais forte o contexto semântico, mais fácil é a integração de uma palavra e menor o sinal N400, que indica o erro semântico. Neste sentido, eles sustentam que a prosódia emocional guia a contextualização de palavras similares (SCHIRMER; KOTZ; FRIEDERICI, 2002).

Em outro teste, Annett Schirmer e outros verificaram que a prosódia auxilia na seleção de alternativas de contexto semântico. Esta constatação foi baseada na ativação cerebral realizada na busca da valência (conjunto de relacionamentos) de uma palavra com e sem uso da prosódia com entonações emocionais (SCHIRMER et al., 2004; SCHIRMER; KOTZ; FRIEDERICI, 2005).

3.2 Em busca de um modelo computacional

Após a constatação da influência da prosódia em três etapas do processamento da linguagem falada, aqui é proposta uma etapa a mais no modelo de Friederici. Em paralelo às fases do MNPAF, propõe-se mais uma fase que permeia todas as demais, iniciando na fase 0 por volta dos 40 ms e terminando na fase 3 em torno dos 800 ms, em valores aproximados, com base nos dados apresentados neste capítulo.

Essa visão ampliada do MNPAF foi ilustrada na figura 3.4, que mostra como a prosódia não pode ser ignorada quando da análise da linguagem falada. Aliando-se o modelo neurocognitivo à proposta do desenvolvimento de um sistema computacional coerente com o processamento do MNPAF, torna-se necessário o levantamento de tecnologias que possam sustentar este novo modelo computacional.

As tecnologias cabíveis são investigadas nos próximos dois capítulos, relativos à análise do sinal de fala (cap. 4) e processamento da linguagem (cap. 5). Acredita-se que, juntando-se a computação destas duas grandes áreas, a exemplo dos sistemas existentes, será possível a proposição de um modelo computacional compatível com o modelo neurocognitivo.

4 ANÁLISE DO SINAL DE FALA

O sinal de fala recebido pelo computador é submetido a diversas transformações até estar adequado ao processamento da linguagem. Isso porque a fala é oriunda da modulação do ar que sai dos pulmões. Esta modulação inicia com as cordas vocais, que propiciam a onda fundamental (F0) do sinal da fala. A F0 é refletida no trato vocal, que vai da faringe até as cavidades nasais, e ainda modulada pela língua e os lábios. Como resultado, tem-se uma onda complexa, com inúmeras informações acerca do locutor e da locução.

Na presente tese o processamento do sinal é focado na locução. As técnicas apresentadas neste capítulo são referentes, portanto, à extração de informações relativas à locução.

Nas seções a seguir, apresenta-se a forma de recepção física do sinal e o processamento que normalmente é realizado para a identificação da parte da voz referente aos conteúdos lexicais. Em outras palavras, preocupa-se aqui com a comunicação através da linguagem falada.

Após identificado o sinal contendo a linguagem, passamos a extrair o conteúdo que identifique os itens lexicais, em especial para representação de palavras. Para tanto, são aqui expostas duas técnicas tradicionais: uma que utiliza a transformada de Fourier e outra que aplica a predição linear.

Entretanto, nesta tese é utilizada a representação do sinal lingüístico através de transformadas ondeletas (wavelets). Seu uso é justificado por sua capacidade representação tempo-freqüência, que supera o potencial de localização apenas em freqüência dos métodos tradicionais.

Ainda outra propriedade foi explorada nas transformadas ondeletas, que é a possibilidade de estimação da F0, o que permite a análise da prosódia (entonação) da fala. Esta propriedade será posteriormente explorada no modelo computacional proposto no capítulo 6.

4.1 Recepção e pré-processamento

O processamento da onda sonora da fala inicia na transformação desta em onda elétrica através de um transdutor como um microfone, por exemplo. O microfone possui uma placa sensora que, ao vibrar com as ondas sonoras, gera tensão elétrica em uma pequena bobina. Esta onda elétrica chega a um conversor analógico-digital que a transforma, por sua vez, em códigos binários. A cada instante de tempo t é feita uma medida da tensão da corrente (equivalente à amplitude da onda elétrica), a qual é codificada em binário. Esta medida realizada num tempo t é chamada de *amostra*.

As amostras devem ser capturadas numa taxa que seja no mínimo o dobro da freqüência máxima de oscilação da onda a processar, também chamada de *freqüência de Nyquist*

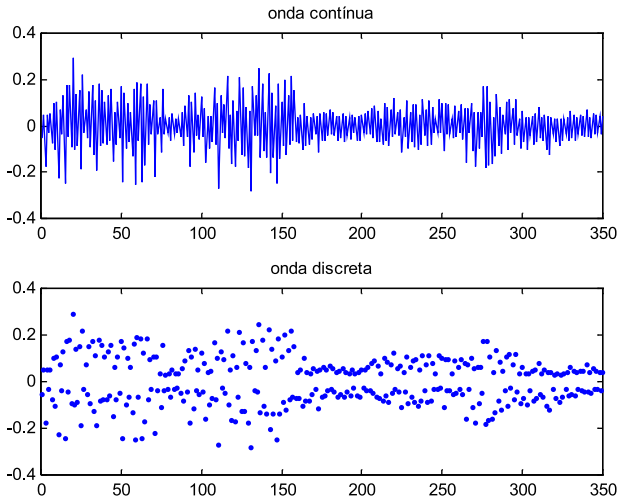

Figura 4.1: Onda contínua e onda discretizada.

(RABINER; SCHAFER, 1978). Caso a amostragem seja abaixo do dobro da freqüência de Nyquist, ocorre o efeito de sobreposição no espectro de freqüência (*aliasing*). Caso ocorra a sobreposição das ondas amostradas no domínio freqüência, isso significa que possui dados insuficientes para ser reconstruída como onda analógica. Desta forma, fazendo a amostragem acima do dobro da freqüência de Nyquist, garante-se que a onda discretizada (figura 4.1b) nas amostras corresponde fielmente à onda real contínua (figura 4.1a).

Desta forma, para processamento do sinal da fala, sabe-se que a freqüência das ondas da fala não passam de 4KHz e portanto uma amostragem acima de 8KHz seria adequada. Atualmente, a taxa mínima de amostragem de sinais de fala estabelecida em sistemas computacionais é de 11KHz, suficiente, portanto, para o processamento de fala a um nível básico.

Após o sinal capturado e gravado em meio digital, há a necessidade de normalizá-lo, uma vez que as gravações podem estar em diferentes intensidades de sinal, muitas vezes causadas pela distância do microfone à boca. Para padronizar as diferentes intensidades, estipula-se uma faixa de variação de amplitude *va*, que é multiplicado pela amostra *A* e nivelado pelo maior sinal amostrado *Amax* , na forma:

$$N_i = \frac{A_i * va}{Amax} \tag{4.1}$$

As amostras normalizadas *N* estarão dentro de uma faixa de amplitude estipulada, em geral em torno de zero, como por exemplo *va* = 0,5, uma vez que a tensão elétrica gira entre -0,5V e 0,5V (VALIATI, 2000).

Uma vez normalizadas, o trecho falado nas amostras deve ser identificado. Para tanto, utilizam-se técnicas para detecção do início e fim da locução. Esta detecção é feita com o cálculo da energia do sinal e do número de cruzamentos por zero (RABINER; SCHAFER, 1978). Quando a energia cresce, significa que houve o início da locução, quando declina, significa que finalizou a locução. Porém, há algumas palavras nas quais é sutil a percepção do crescimento da energia, para tanto, utiliza-se também a técnica do número de cruzamentos por zero da onda. Quanto mais vezes a onda cruzar o eixo zero num dado

período de tempo, mais probabilidade há que esteja ocorrendo uma locução.

Assim, seja *x* o vetor de amostras de sinal dentro de determinado período de tempo e *a* uma taxa de refinamento que varia de 0 a 1, tem-se o cálculo da energia como sendo (RABINER; SCHAFER, 1978):

$$E_n = aE_{n-1} + x^2(n) \tag{4.2}$$

Mas como o cálculo de energia pode variar muito de uma amostra *n* para outra, calcula-se apenas a magnitude média (RABINER; SCHAFER, 1978):

$$M_n = aM_{n-1} + |x(n)| \tag{4.3}$$

Estipula-se, então, um limiar para identificação do início e fim da locução. Em geral, este limiar é verificado experimentalmente conforme as gravações das locuções que se pretende processar.

Para verificar o número de cruzamentos por zero, necessita-se contabilizar a diferença entre duas amostras, primeiramente marcando se a amostra é positiva ou negativa:

$$y(n) = \begin{cases} 1, & se\ x(n) \geq 0 \\ -1, & se\ x(n) < 0 \end{cases} \tag{4.4}$$

Desta forma, faz-se a contagem para determinado período de tempo *t*:

$$Z_m = \sum_{i=1}^{t} y(i) \tag{4.5}$$

A partir da contagem da variação do cruzamento por zero em *m* instantes de tempo, estipula-se novamente um limiar para consideração do número mínimo de cruzamentos que indicaria início da locução.

Uma vez computados os dados, pode-se fazer a comparação com os valores obtidos no processamento da energia no mesmo período e proceder a identificação dos pontos de início e de fim da locução. Com base nestes pontos, apenas as amostras no intervalo definido continuarão a ser processadas.

4.2 Método da extração de coeficientes cepstrais

O sinal discreto da locução, após ter sido devidamente processado, serve como base para a extração das características que permitirão a sua identificação por sistemas classificatórios do sinal de fala. Há duas técnicas amplamente difundidas para extração de coeficientes cepstrais: transformada de Fourier e codificação por predição linear (*Linear Predictive Coding* - LPC).

A técnica da extração do *cepstrum* baseia-se na idéia de que há sinais superpostos, formando um sinal só, composto. Estes sinais combinados são chamados de *sistemas homomórficos*, por vários sinais diferentes comporem uma só forma. A superposição de sinais é matematicamente conhecida como *convolução* e obedece à formulação geral (RABINER; SCHAFER, 1978):

$$y(n) = \sum_{k=-\infty}^{\infty} x(k)\, h(n-k) = h(n) \star x(n) \tag{4.6}$$

onde *x* e *h* são os sinais superpostos (convoluídos).

No que se refere ao sinal de fala, acredita-se que haja uma superposição de sinais do impulso de resposta do trato vocal e da radiação (movimento dos lábios), bem como o pulso glotal (movimento das cordas vocais) e a excitação (fluxo de ar dos pulmões) (RABINER; SCHAFER, 1978). Para separação destes sinais, utiliza-se a deconvolução homomórfica, também chamada de extração do cepstrum.

Para este processo de deconvolução, é necessário que se converta o sinal do domínio tempo (amplitude do sinal x tempo) para o domínio freqüência (amplitude da freqüência x freqüência). Para tanto, é necessária a aplicação de uma transformada do tipo Z, que faz a transposição do plano cartesiano para o plano radial. A transformada do tipo Z mais utilizada no processamento de fala é a transformada de Fourier.

4.2.1 Codificação pela transformada de Fourier

A transformada de Fourier é a decomposição de uma onda através de cossenóides e senóides. As cossenóides formadas representam a amplitude (magnitude) e as senóides seriam a fase da onda decomposta. O cálculo das decomposições da magnitude e da fase é feito pelas equações (SMITH, 1999):

$$M(k) = \sum_{i=0}^{N-1} x(i)\, cos(2\pi ki/N) \tag{4.7}$$

e

$$F(k) = -\sum_{i=0}^{N-1} x(i)\, sen(2\pi ki/N) \tag{4.8}$$

onde *N* seria o número de amostras de uma janela *k* de tempo. Esta janela de tempo pode ser fixa ou variável conforme o tempo total. Para o preenchimento de todo o tempo da locução são necessárias, portanto, diversas janelas, ou seja, o *janelamento* do sinal em processamento.

As janelas utilizadas neste processo devem ser sobrepostas para realizar uma sincronia temporal, uma vez que a aplicação da transformada de Fourier não faz nenhuma correspondência do domínio tempo. Para que seja possível uma correspondência em termos de tempo, as janelas devem sobrepor-se no mínimo em 50% para janelas retangulares e 75% para outras (VALIATI, 2000).

As janelas retangulares são aquelas onde simplesmente utilizam-se os valores de amostra dentro do intervalo da janela. As demais são *filtragens*, ou seja, convoluções das amostras com uma equação de filtragem que proporciona decaimento do sinal nas margens da janela. O objetivo deste decaimento é reduzir a distorção da onda causada pelo janelamento.

Um exemplo de filtro de janelamento é a chamada *janela Hamming* (RABINER; SCHAFER, 1978):

$$h(n) = \begin{cases} 0,54 - 0,46\, cos(2\pi n/(N-1)), & 0 \leq n \leq N-1 \\ 0 & caso\ contrário \end{cases} \tag{4.9}$$

onde *N* é o número de amostras da janela e *n* a amostra calculada.

Uma vez escolhido o método para o janelamento, calcula-se a transformada de Fourier

$$X(k) = \Im\{x(n) \star h(n)\} \tag{4.10}$$

e, para obtenção do cepstrum realiza-se a deconvolução homomórfica, operação inversa à convolução (RABINER; SCHAFER, 1978). Assim, aplica-se o logaritmo sobre a magnitude

$$\widehat{X}(k) = log\,|X(k)| \tag{4.11}$$

ou, para efeitos práticos, calcula-se somente a magnitude e seu logaritmo, conforme analisada anteriormente

$$\widehat{M}(k) = log(M(k)) \tag{4.12}$$

Caso seja necessária a reconstrução do sinal, aplica-se a transformada inversa:

$$\widehat{x}(n) = \Im^{-1}\left\{\widehat{X}(k)\right\} \tag{4.13}$$

ou seja (RABINER; SCHAFER, 1978; SMITH, 1999):

$$\widehat{x}(n) = \sum_{k=0}^{N/2} \widehat{M}(k)\, cos(2\pi ki/N) + \sum_{k=0}^{N/2} F(k)\, sen(2\pi ki/N) \tag{4.14}$$

Para a representação da locução com base no cepstrum calculado, tem-se interesse nos valores dos 10 a 20 primeiros elementos de $\widehat{x}$ (VALIATI, 2000), o que deve corresponder ao sinal da glote e trato vocal, como comentado anteriormente.

4.2.1.1 Transformada de Fourier e a escala mel

Na tentativa de aproximar a resposta de freqüência dos coeficientes gerados à capacidade de percepção do sistema auditório humano, foi utilizada a escala de freqüências denominada *mel*. Esta escala é linear abaixo de 1KHz e logarítmica acima desta faixa. Diferentes equações têm sido apresentadas para representar esta escala, tais como (GODINO-LLORENTE; GÓMEZ-VILDA, 2004):

$$f_{mel}(x) = \frac{1000}{\log 2}\left(1 + \frac{f(x)}{1000}\right) \tag{4.15}$$

onde $f(x)$ é a freqüência real da amostra x, e como

$$f_{mel}(x) = 2595 \cdot \log\left(1 + \frac{f(x)}{700}\right) \tag{4.16}$$

que foi aplicada a um contexto de normalização de trato vocal (PITZ; NEY, 2005).

A aplicação da escala mel na extração de coeficientes cepstrais (MFCC - Mel Frequency Cepstral Coefficients), segundo (PITZ; NEY, 2005), aplica a equação 4.10 e após calcula a potência espectral por

$$S(k) = |X(k)|^2 \tag{4.17}$$

e a energia S de cada quadro mel i dentro da janela de análise com

$$S_i = W_i(k) \cdot S(k) \tag{4.18}$$

onde $1 \leq i \leq M$ e M é o número de quadros na escala mel, que variam entre 20 e 24. A cada quadro i está associada uma função triangular Wna escala mel.

A extração dos coeficientes é obtida através da aplicação da derivada discreta do cosseno (PITZ; NEY, 2005):

$$c_{mel} = \sum_{i=1}^{M} \log(S)\, \cos\left[m\, i - 0,5\, \frac{\pi}{M}\right] \tag{4.19}$$

onde $1 \leq i \leq L$ e L é a ordem desejada do MFCC.

4.2.2 Codificação por Predição Linear

A predição linear é o método que modela o trato vocal como um filtro linear do sinal proveniente da excitação, do tipo (RABINER; SCHAFER, 1978; SAYOOD, 1996):

$$y(n) = \sum_{k=1}^{p} a(k)\, y(n-k) + e(n) \tag{4.20}$$

onde *a(k)* são os coeficientes do filtro e *e(n)* é o erro (diferença) existente entre um trecho de *p* valores *y* anteriores e o valor *y* atual.

Desta forma, pode-se estimar o próximo valor com base numa combinação linear dos valores anteriores, calculando-se o *y* futuro (RABINER; SCHAFER, 1978):

$$\widehat{y}(n) = \sum_{k=1}^{p} a(k)\, y(n-k) \tag{4.21}$$

e o erro

$$e(n) = y(n) - \widehat{y}(n) \tag{4.22}$$

A codificação por predição linear é o processo de cálculo dos coeficientes a que permitirão a estimativa do sinal futuro. Para tanto, calculam-se aproximações dos valores de erro através do *método da autocorrelação*, que calcula o erro médio quadrado mínimo, e o *método da covariância*, que calcula o erro médio quadrado.

O erro médio por autocorrelação é calculado, dentro de uma janela *j* de tempo, na forma (RABINER; SCHAFER, 1978):

$$E(j) = R(j)(0) - \sum_{k=1}^{p} a(k)\, R(j)(k) \tag{4.23}$$

sendo que (RABINER; SCHAFER, 1978; OLIVEIRA, 1998):

$$R(j)(k) = \frac{1}{N-k} \sum_{n=0}^{N-1-k} y(j)(n)\, y(j)(n+k) \tag{4.24}$$

onde *N* é o tamanho da janela *j* e *k* o elemento que está sendo calculado. Estas equações podem ser resolvidas por (RABINER; SCHAFER, 1978; SAYOOD, 1996; OLIVEIRA, 1998):

$$\sum_{k=1}^{p} a(k)\, R(j)(|i-k|) = R(j)(i) \tag{4.25}$$

o que representa o sistema para uma janela *j* de *p* valores:

$$\begin{bmatrix} a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(p) \end{bmatrix} \begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \tag{4.26}$$

Este sistema pode ser resolvido com o algoritmo Levinson-Durbin, que realiza o cálculo de constantes chamadas coeficientes de reflexão ou correlação parcial (PARCOR - *parcial correlation*). O algoritmo segue os passos (SAYOOD, 1996):

1. Inicialize com zero o vetor de erro médio E, a matriz R e o índice *i* que corresponde à ordem do filtro;
2. incremente i em 1;
3. calcule $k(i) = \left(\sum_{j=1}^{i-1} a(j)^{(i-1)}\, R(i-j+1) - R(i)\right) \Big/ E(i-1)$;
4. inicialize $a(i)^i = k(i)$;
5. calcule $a(j)^{(i)} = a(j)^{(i-1)} + k(i)\, a(i-j)^{(i-1)}$, para *j* = 1, 2, ..., *i*-1;
6. calcule $E(i) = (1 - k(i)^2)\, E(i-1)$;
7. Se $i < p$ volte ao passo 2.

Assim, são obtidos os coeficientes a no método de autocorrelação, que serão passíveis de uso para posterior identificação da locução.

Já para o método de covariância, é usado o cálculo do erro médio quadrado para uma janela *j* de tempo, cujo tamanho é *N* (RABINER; SCHAFER, 1978):

$$E(j) = \sum_{k=0}^{N-1} e(j)(k)^2 \tag{4.27}$$

que pode ser descrita da forma (RABINER; SCHAFER, 1978; OLIVEIRA, 1998):

$$\phi(j)(i,k) = \frac{1}{N} \sum_{m=0}^{N-1} y(j)(m-i)\, y(j)(m-k) \qquad \begin{matrix} 1 \leq i \leq p \\ 0 \leq k \leq p \end{matrix} \tag{4.28}$$

Esta fórmula pode ser satisfeita, de forma semelhante à autocorrelação, pela equação (RABINER; SCHAFER, 1978):

$$\sum_{k=1}^{p} a(k)\, \phi(j)(i,k) = \phi(j)(i,0) \qquad i = 1, 2, ..., p \tag{4.29}$$

que corresponde ao sistema para uma janela *j* de *p* valores:

$$\begin{bmatrix} a(1) \\ a(2) \\ a(3) \\ \vdots \\ a(p) \end{bmatrix} \begin{bmatrix} \phi(1,1) & \phi(1,2) & \phi(1,3) & \cdots & \phi(1,p) \\ \phi(2,1) & \phi(2,2) & \phi(2,3) & \cdots & \phi(2,p) \\ \phi(3,1) & \phi(3,2) & \phi(3,3) & \cdots & \phi(3,p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi(p,1) & \phi(p,2) & \phi(p,3) & \cdots & \phi(p,p) \end{bmatrix} = \begin{bmatrix} \phi(1,0) \\ \phi(2,0) \\ \phi(3,0) \\ \vdots \\ \phi(p,0) \end{bmatrix} \tag{4.30}$$

Este sistema matricial $\alpha\Phi = \psi$ (matrizes de coeficientes a, $\phi(i,j)$ e $\phi(i,0)$, respectivamente) pode ser resolvido através da decomposição Cholesky, que transforma a matriz simétrica Φ em duas matrizes diagonais e uma simétrica na forma $\Phi = VDV^t$, onde V é uma matriz triangular, V^t sua transposta e D uma matriz diagonal, calculados na forma:

$$v(i,j)\, d(j) = \phi(i,j) - \sum_{k=1}^{j-1} v(i,k)\, d(k)\, v(j,k) \qquad 1 \leq j \leq i-1 \tag{4.31}$$

e os elementos diagonais:

$$d(i) = \phi(i,i) - \sum_{k=1}^{i-1} v(i,k)^2\, d(k) \qquad i \geq 2 \tag{4.32}$$

Uma vez resolvido Φ, passa-se à resolução de α. Sabendo que $VDV^t\alpha = \psi$, pode-se definir que $Y = DV^t\alpha$ e atribuir $VY = \psi$ e calcular Y. A partir daí resolve-se α fazendo $V^t\alpha = D^{-1}Y$, de onde serão obtidos os coeficientes para posterior classificação de padrões de fala.

4.3 Método da Transformada Ondeletas

4.3.1 Da Transformada de Fourier à Transformada Ondeletas

As ondeletas fazem decomposições como a transformada de Fourier, mas em função de dois parâmetros: escala (também chamada dilatação) e deslocamento temporal (ou translação). Isso é possível porque as ondeletas utilizam, ao invés de apenas senos e cossenos, funções que permitem a localização de sinais tanto no domínio freqüência quanto no tempo.

As diferenças entre as transformadas podem ser apresentadas por suas séries. A transformada contínua de Fourier é representada em função de senos e cossenos na forma (KAHANE; LEMARIÉ-RIEUSSET, 1995; MORETTIN, 1999):

$$\widetilde{f}(x) = \sum c_n e^{inx} \tag{4.33}$$

ou seja,

$$f(x) = \frac{a_0}{2} \sum_1 (a_n^\infty \cos nx + b_n \sin nx) \tag{4.34}$$

que tem

$$c_n = \int f(x) e^{-inx} \frac{dx}{2\pi} \tag{4.35}$$

ou

$$a_n = \frac{1}{\pi} \int f(x) \cos nx \, dx \tag{4.36}$$

e

$$b_n = \frac{1}{\pi} \int f(x) \sin nx \, dx \tag{4.37}$$

onde $f(x)$ é a função a ser transformada. As funções 4.36 e 4.37 são representações contínuas das formas discretas 4.7 e 4.8, respectivamente.

Segundo Gomes, os coeficientes c_n da eq. 4.35 medem a amplitude da freqüência do componente n (GOMES; VELHO; GOLDENSTEIN, 1997). Por isso é dito que a transformada de Fourier é uma *análise de domínio freqüência*. O comportamento desta transformada em relação ao tempo pode ser visualizada na figura 4.2 a), onde se percebe que há diferentes níveis de freqüência para o mesmo instante de tempo.

Analogamente, a transformada contínua por ondeletas é representada na forma (DAUBECHIES, 1992; GOMES; VELHO; GOLDENSTEIN, 1997; MORETTIN, 1999):

$$\widetilde{f}(x) = \int f(x) \psi_{a,b}(x) \, dx \tag{4.38}$$

e

$$\psi_{a,b}(x) = \frac{1}{\sqrt{|a|}} \psi \left(\frac{x-b}{a} \right) \tag{4.39}$$

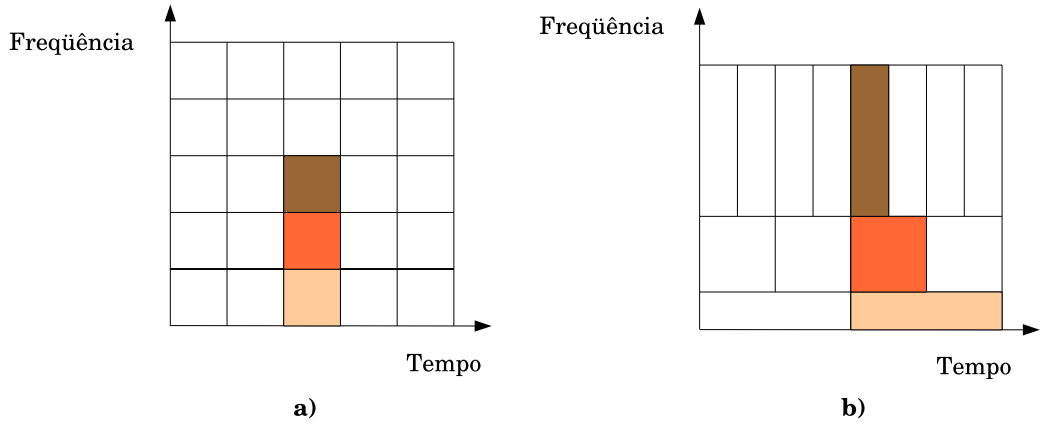

Figura 4.2: Comparação entre a divisão do tempo com relação à freqüência em Fourier (a) e por ondeletas (b).

onde $f(x)$ é a função a ser transformada, a é o parâmetro de controle de escala e b de deslocamento. Um pequeno valor a implica em escalas muito pequenas para as altas freqüências e o valor b sempre vai auxiliar na localização no tempo, ou seja, será próximo a x. Por isso que a transformada ondeletas é chamada de *análise de domínio tempo-freqüência*. Na figura 4.2 b), pode-se observar que há diferentes freqüências para diferentes escalas de tempo.

Uma das ondeletas mais antigas foi desenvolvida por Haar em 1909, a qual possui a ondeleta-mãe definida por

$$\psi(x) = \begin{cases} 1 & se\ x \in [0, \frac{1}{2}) \\ -1 & se\ x \in [\frac{1}{2}, 1) \\ 0 & se\ x < 0\ ou\ x > 1 \end{cases} \tag{4.40}$$

cujos efeitos de escala e deslocamento dos índices podem ser observados na figura 4.3.

A função da eq. 4.39 é chamada de ondeleta-mãe, uma vez que é a responsável pela transformada. Esta função tem como uma das condições ser ortogonal, para possibilitar a correta reconstrução do sinal pela transformada inversa.

Para toda transformada existe a operação inversa, que é chamada de reconstrução. A decomposição, como as transformadas anteriormente apresentadas, gera coeficientes em forma de funções trigonométricas. A reconstrução, por sua vez, recupera a onda original a partir dos coeficientes da transformada. Como a transformada inversa não é relevante na presente Tese, ela não será aqui desenvolvida.

As transformadas contínuas geram uma quantidade impraticável de cálculos, limitada apenas à capacidade de representação numérica. Para sistemas computacionais, é portanto necessário o uso de transformadas discretas, ou seja, com limitação de intervalos de tempo.

Na transformada de Fourier a única alternativa para a associação com o domínio tempo é o janelamento, realizado através da fixação de pequenos períodos de tempo para determinação dos coeficientes, calculados agora na forma de uma função de translação *g* aplicada à eq. 4.35:

$$c_n^t = \int g(x-t) f(x) e^{-inx} \frac{dx}{2\pi} \tag{4.41}$$

Na transformada ondeletas bastam as atribuições, na eq. 4.39, de $a = a_0^m$ e $b = nb_0a_0^m$.

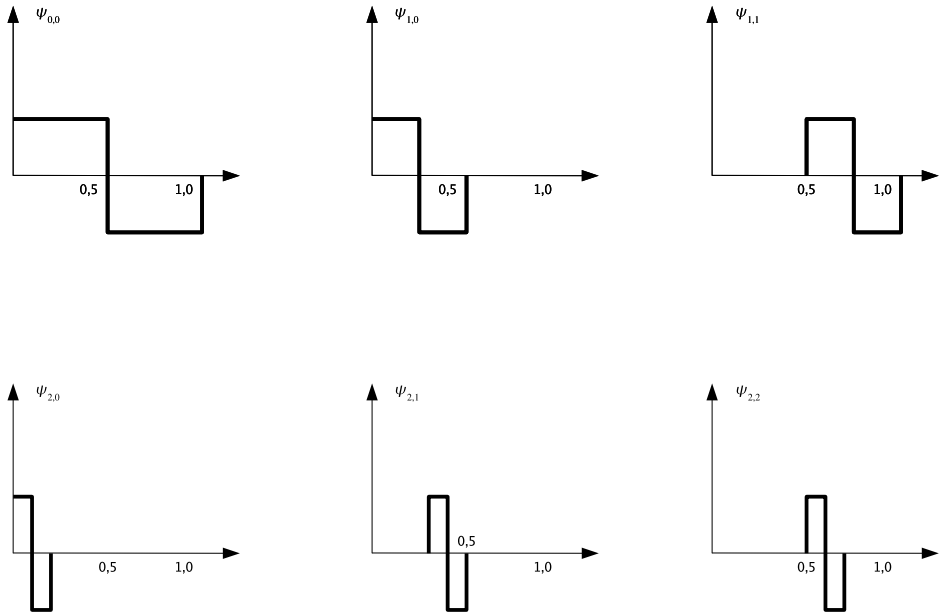

Figura 4.3: Efeitos de escala (seqüência superior) e deslocamento na ondeleta Haar.

Desta forma, *a* (escala) e *b* (deslocamento) são valores constantes e *m* e *n* são as variações de dilatação (largura da janela de tempo) e translação (tamanho do deslocamento da janela) da transformada.

Assim, tem-se a transformada discreta com a ondeleta-mãe:

$$\psi_{m,n}(x) = \frac{1}{\sqrt{a_0^m}}\psi\left(\frac{x - nb_0a_0^m}{a_0^m}\right) \tag{4.42}$$

As ondeletas-mãe, como a da eq. 4.42, são funções de *suporte compacto*, pois possuem uma banda limitada a determinado intervalo. Dito de outra forma, elas funcionam como um filtro passa-banda (GOMES; VELHO; GOLDENSTEIN, 1997).

4.3.2 Análise de Multiresolução da Transformada Ondeletas

Como as ondeletas podem funcionar como filtros, é possível a construção de bancos de filtros através delas, possibilitando uma análise em multiresolução. Este tipo de análise pode ser comparada com um ajuste de escala para aproximação e afastamento (*zoom-in / zoom-out*) de uma imagem. Isso significa que, aplicando-se uma função de escala à ondeleta-mãe é possível a definição de coeficientes mais detalhados do sinal.

Neste ponto de vista, Ingrid Daubechies observa que a transformada de Fourier pode ser considerada uma análise *grossa* do sinal em estudo, enquanto as ondeletas permitem uma análise refinada deste mesmo sinal. A partir disso, Stèphane Mallat e Yves Meyer proporcionaram, em 1986, uma base teórica para análise multiresolução em ondeletas (DAUBECHIES, 1992).

Assim, define-se a função escala:

$$\phi(x) = \sqrt{2}\sum_k h_k\phi(2x - k) \tag{4.43}$$

onde h_k é um filtro passa-baixa e *k* o índice de escala. A ondeleta-mãe será colocada nesta

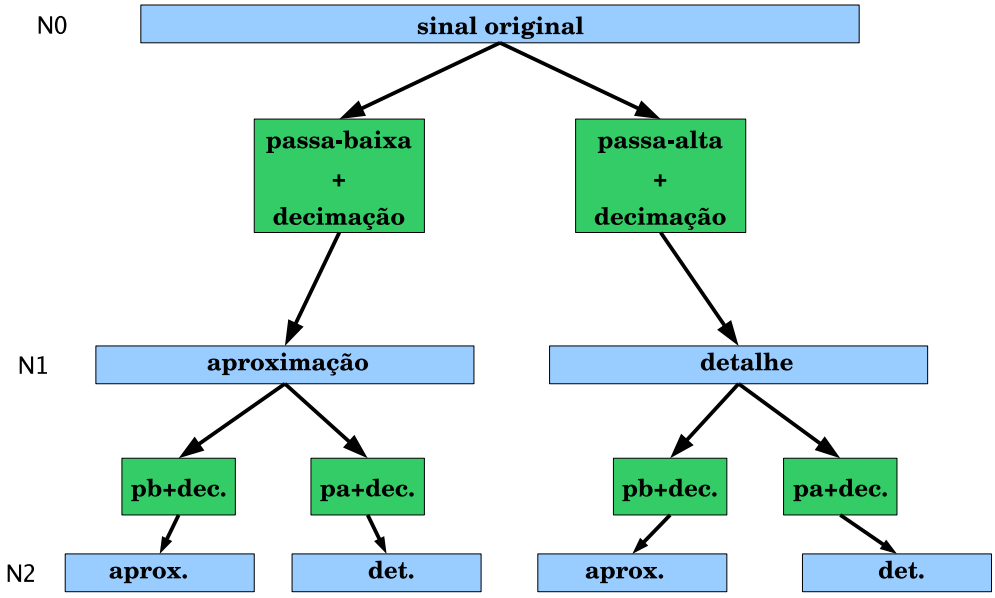

Figura 4.4: Árvore de análise multiresolução de ondeletas.

escala por:

$$\psi(x) = \sqrt{2}\sum_{k} g_k \psi(2x - k) \tag{4.44}$$

onde g_k é um filtro passa-alta.

A função de escala da eq. 4.43 gera coeficientes aproximados, enquanto a ondeleta-mãe da eq. 4.44 gera coeficientes detalhados do sinal. A cada aplicação de filtro também é realizada uma decimação diádica, que é a retirada de um ponto de amostragem para cada par de pontos do sinal filtrado. A aplicação sucessiva destas funções proporciona uma árvore binária com coeficientes cada vez mais refinados, como ilustra a figura 4.4.

4.3.3 Uso de ondeletas para extração de características do sinal de fala

4.3.3.1 Modelo de processamento da audição

Segundo (YANG; WANG; SHAMMA, 1992), os modelos computacionais de representação da audição podem ser compreendidos em três estágios: análise, transdução e redução. Na fase de análise, a representação dá-se como uma visão funcional da cóclea, a qual pode ser vista como um banco de filtros passa-banda em paralelo. Os filtros de análise de freqüências, nesta representação, acima de cerca de 800 Hz caracterizam-se por diferenciarem-se em escala. Por isso diz-se que há uma resposta de freqüência logarítmica na referida faixa de análise. Já na faixa abaixo de 800 Hz o comportamento é mais linear, principalmente abaixo de 500 Hz. A resposta do sinal desta fase de análise é, portanto, uma convolução (filtragem) como a apresentada anteriormente na equação 4.6 análoga à realizada pelas ondeletas (YANG; WANG; SHAMMA, 1992):

$$y_1(s;t) = h(t;s) \star_t x(t) \tag{4.45}$$

onde $h(t;s)$ é a filtragem correspondente à cóclea no local s com entrada $x(t)$ e o operador $\star_t$ indica a convolução no tempo t.

O segundo estágio, utilizado em algoritmos de segmentação de fala, corresponde à transdução das ondas sonoras em atividade neuronal. Na cóclea há células com cílios que

geram potenciais elétricos que são emitidos ao sistema de audição central. Este processo pode ser representado por

$$y_2(t;s) = g\left(\frac{\partial y_1(t;s)}{\partial_t}\right) \star_t w(t) \tag{4.46}$$

onde w é um filtro passa-baixa para suavização da janela de tempo e a função sigmóide g é na forma

$$g(u) = \frac{1}{1+^{-\gamma u}} - \frac{1}{2} \tag{4.47}$$

onde γ é uma constante de ganho da função.

Ao chegar neste sistema, ocorre a fase de redução, uma vez que seus neurônios trabalham numa velocidade menor e avaliam os sinais resultantes da inibição lateral das células auditivas. Como conseqüência, ocorre uma interação lateral dos impulsos na forma

$$y_3 = \left(g'\left(\frac{\partial y_1(t;s)}{\partial_t}\right) \cdot \left(\frac{\partial \partial y_1(t;s)}{\partial_t \partial_s}\right)\right) \star_t w(t) \star_s v(s) \tag{4.48}$$

onde tem-se a derivada tanto em função do tempo t como em função da janela s, g'é a derivada da função 4.47, w é a filtragem em função do tempo e o operador de convolução $\star_s$ é aplicado sobre o filtro v da janela. Segundo (YANG; WANG; SHAMMA, 1992), a função g' é calculada como uma função de passos Heaviside, que é aproximada por uma função Dirac delta para integração dos impulsos semelhante à gaussiana (SHAMMA; MORRISH, 1987):

$$g'(x) \approx \delta_\alpha(x) \equiv \frac{1}{\pi}\,\frac{\alpha}{x^2+\alpha^2} \tag{4.49}$$

e as derivadas parciais podem ser simplificadas na forma (YANG; WANG; SHAMMA, 1992):

$$\frac{\partial y_1(t;s)}{\partial_t} \approx y_1(i;k)$$

e

$$\frac{\partial \partial y_1(t;s)}{\partial_t \partial_s} \approx \frac{\partial y_1(t;k+1)}{\partial_t} - \frac{\partial y_1(t;k)}{\partial_t}$$

ou seja

$$\frac{\partial \partial y_1(t;s)}{\partial_t \partial_s} \approx y_1(i+1;k+1) - y_1(i;k+1)$$

com $t = i$ para o k-ésimo canal de análise.

Após a filtragem do sinal, o resultante é recomposto:

$$Y_4(t;s) = \max(y_3(t;s), 0) \quad . \tag{4.50}$$

Por fim, o sinal filtrado e recomposto é integrado em janelas de análise (espectral) que giram entre 10 e 20 ms (YANG; WANG; SHAMMA, 1992):

$$y_5(t;s) = y_4(t;s) \star_t \Pi_T(t) \tag{4.51}$$

onde Π_T é um filtro para uma janela de duração T. Em termos práticos, calcula-se:

$$y_5(t;s) \approx \frac{1}{T}\sum_i \max\left(\frac{\partial \partial y_1(t_{i;s};s)}{\partial_t \partial_s}, 0\right)$$

Segundo (YANG; WANG; SHAMMA, 1992), este último estágio de redução pode ser realizado pela transformada de Fourier, uma vez que seu comportamento é análogo ao resultado obtido quanto à análise espectral.

4.3.3.2 As ondeletas e a escala mel

Como analisado na seção 4.2.1, para que o processamento por Fourier tenha uma resposta de freqüência logarítmica, tal como a audição humana, foi necessária a aplicação da escala mel e derivada do cosseno, na forma do algoritmo MFCC. A transformada ondeletas, por outro lado, já possui uma resposta logarítmica (GOLDSTEIN, 1994), o que a torna mais adequada para extração das características da fala. Por suas características, as ondeletas também são utilizadas em sistemas biônicos de implantes de cóclea (YAO; ZHANG, 2002).

Um problema detectado no algoritmo MFCC é o fato da derivada do cosseno criar apenas uma representação para todas as faixas de freqüência, e basta uma das faixas estar corrompida por ruído para afetar todos os coeficientes gerados (TUFEKCI; GOWDY, 2000). Além disso, com a aplicação do MFCC pode ocorrer de um quadro conter mais de um fonema, pois o espectro do fonema de alta freqüência pode dominar sobre outro de baixa freqüência, ocorrendo uma sobreposição (TUFEKCI; GOWDY, 2000). A solução para esses problemas apresentados estaria na análise do sinal por sub-bandas, permitindo a separação das características de cada faixa do sinal, o que é possível através da transformada ondeletas.

Segundo (GUPTA; GILBERT, 2001), é preferível utilizar os coeficientes de ondeletas diretamente do que a energia das sub-bandas, uma vez que não perde a dimensão temporal durante o processamento. Outras pesquisas porém, utilizam a abordagem inversa, como (KIM; YOUN; LEE, 2000; FAROOQ; DATTA, 2004; RICOTTI, 2005), determinando a densidade de potência espectral como característica para identificação de fonemas. A densidade espectral é o cálculo da potência espectral utilizada no algoritmo MFCC (eq. 4.17) dividido pelo número de coeficientes resultantes da filtragem por ondeletas. O resultado deste processo é chamado *envelope espectral*, uma vez que apresenta o contorno da variação da energia dos coeficientes.

Aparentemente, esta última forma de processamento afina-se com a teoria de Yang apresentada na seção anterior, uma vez que propõe-se a extração de coeficientes de ondeletas, resultando na representação de diferentes faixas de freqüência. Posteriormente, realiza-se a obtenção dos dados espectrais a partir dos coeficientes gerados nas diferentes faixas.

A forma de geração dos coeficientes possui basicamente duas abordagens: direta, através de ondeletas-mãe específicas (KADAMBE; SRINIVASAN, 1997; RICOTTI, 2005) e pacotes de ondeletas (CARNERO; DRYGAJLO, 1999; GUPTA; GILBERT, 2001; AVCI; TURKOGLU; POYRAZ, 2005), também conhecida como análise de multiresolução (KIM; YOUN; LEE, 2000; TUFEKCI; GOWDY, 2000), comentada na seção 4.3.2. A aplicação de equações específicas de ondeletas-mãe necessitam maior controle sobre a escala e o deslocamento, enquanto a análise de multiresolução são transformações diádicas, reduzindo a escala sempre pela metade. Em geral, a análise de multiresolução é diretamente representada como um banco de filtros.

4.4 Prosódia

Há determinadas características do sinal de fala que definem os limites de início e fim de uma frase, sua entonação principal e a acentuação das palavras (KOMPE, 1997). Estes fenômenos acústicos que não são representados na linguagem escrita são conhecidos como a **prosódia** da linguagem.

A entonação de uma frase pode mudar seu sentido, como a maneira de dizer *não*,

que pode ter conotação tanto negativa, como afirmativa. A palavra *forma* pode ser falada tanto como *fôrma*, quando se refere a um molde, quanto *fórma*, quando se refere a um feitio, embora seja a mesma grafia. Em suma, a prosódia é essencial para uma adequada compreensão da linguagem falada.

A história da linguagem escrita, aliás, começou pela prosódia. A cultura mesoamericana desenhava objetos para compor palavras sonoras (FERREIRO, 2001). Era como, por exemplo, colocar uma imagem representando um sol seguida de um desenho de um dado para indicar a palavra *soldado*.

A pausa entre as unidades lingüísticas - como as palavras - escritas pelos povos antigos não era feito por espaços, mas por outros símbolos (FERREIRO, 2001). Em determinado momento de sua história, os gregos abandonaram os marcadores e a escrita passou a ser contínua, influenciando outros povos como os romanos. O problema da escrita contínua é que a sonoridade das expressões gráficas deveria ser falado para que o leitor pudesse encontrar o sentido do texto através da audição. O espaço entre as palavras só foi introduzido no século VIII como recurso didático para facilitar a leitura silenciosa de textos (ILLICH, 1995).

Como a pausa, a pontuação dos textos foi também uma evolução histórica. Na Roma Clássica, por exemplo, apenas os leitores especializados sabiam como ler corretamente os textos (FERREIRO, 2001), resolvendo eles mesmos os problemas de interpretação, de forma semelhante ao que hoje é feito no processo computacional de compreensão de fala.

No sentido de resolução dos problemas de compreensão da linguagem falada, a prosódia é um conhecimento imprescindível para uma boa interpretação. De fato, ela pode ser usada como informação em todos os processos analisados no decorrer deste trabalho.

No processamento do sinal de fala, a prosódia pode auxiliar com informações sobre a freqüência fundamental (pitch), que pode variar conforme o locutor (KOMPE, 1997). A partir desse dado, pode-se adequar o processamento do sinal a diferentes níveis de freqüências formantes, melhorando o resultado da extração de coeficientes cepstrais.

No reconhecimento de fala, a prosódia pode ser utilizada para identificação da acentuação da palavra e do tempo entre os fonemas (KOMPE, 1997). A acentuação pode ser importante para o caso da palavra *forma* anteriormente citado e o tempo é importante para definir agrupamentos de palavras e frases.

Na análise sintática, pode se utilizar da prosódia para decidir pontuações de texto, facilitando a desambigüação (KOMPE, 1997). Também nas análises semântica e pragmática a prosódia pode ser utilizada para identificar o ápice de uma entonação frasal (KOMPE, 1997). Em geral, a palavra de maior entonação na frase corresponde à mais importante. Por exemplo, na frase *o **professor** chegou* dá-se ênfase à pessoa que chegou, porém, enfatizando na forma *o professor **chegou***, corresponde ao momento em que a pessoa chegou.

Além de todos estes processos, há o fenômeno paralingüístico dos *acentos enfáticos*, que não foram abordados até o momento neste trabalho. Estes acentos determinam a emoção contida na fala (KOMPE, 1997). Talvez o texto em itálico ou negrito, ou ainda os inovadores *sorrisos*[1] das mensagens eletrônicas (e-mails), sejam formas gráficas de expressar este tipo de acento. Mas como estes acentos não são normalmente registrados na língua escrita, perde-se uma valiosa informação, contida apenas na linguagem falada, não processada pelos métodos analisados anteriormente.

[1]Também chamados *smiles*. São expressões criadas por usuários de correio eletrônico (*e-mail)* para representar seu estado de espírito. Exemplos deles são os caracteres **:-)** (felicidade / bom humor), **:-(** (tristeza), **8-0** (espanto), entre outros.

4.4.1 Elementos prosódicos do sinal de fala

A prosódia, no nível do sinal de fala, possui elementos básicos e compostos. Os elementos básicos são as características do sinal como a energia (loudness), freqüência fundamental (pitch), qualidade vocal, duração e pausa. Os elementos compostos nada mais são do que a ação dos elementos básicos sobre o tempo (KOMPE, 1997).

Os conceitos básicos de **freqüência fundamental** e **energia** foram apresentados no início deste capítulo. A **qualidade vocal** refere-se à estrutura espectral dos fonemas, a partir da qual são investigados efeitos como a *laringealização*, que são períodos da fala com excitação irregular, e a *vogal central*, que é uma vogal que perde sua acentuação por estar sobre o centro do intervalo de freqüência coberto sobre todas as vogais.

A **duração** da fala depende da velocidade com que os fonemas são pronunciados. Como cada fonema possui uma média e desvio padrão próprios a ele, deve ser feita uma normalização dos valores para se obter uma velocidade média de uma palavra. Por outro lado, podem ser construídos modelos para reconhecimento dependentes da duração de cada fonema, visando a previsão de construções de palavras.

A **pausa** pode ser simplesmente um silêncio ou um ruído de fundo que entremeia a fala, o que é chamado de *pausa não-preenchida*. Mas a pausa pode ser também vocalizada, através de expressões como *humm*, ao que é chamada de *pausa preenchida*.

O elementos prosódicos compostos são a entonação, a acentuação, a frase prosódica, o ritmo e a hesitação (KOMPE, 1997). A **entonação** é basicamente a variação da freqüência fundamental. Como os valores da onda fundamental são dependentes do locutor, variando conforme sexo e idade por exemplo, é avaliada apenas o comportamento da envoltória desta onda. Este comportamento é uma informação prosódica que é rotulada para posterior processamento. Exemplos de etiquetas são: descida (*fall*) - F; subida (*rise*) - R; continuação-subida (*continuation-rise*) - CR. Estas etiquetas são extensamente usados nos demais elementos compostos, como analisado na acentuação a seguir.

A **acentuação** refere-se a uma sílaba acentuada dentro de um contexto. A acentuação pode vir da própria palavra onde a sílaba está inserida ou da entonação dada pelo locutor. Em geral, são etiquetados o acento primário ou frasal (PA - *primary* ou *phrase accent*), o acento enfático ou forte (EA - *emphatic* ou *strong accent*) e o acento secundário ou fraco (NA - *secondary* ou *weak accent*). A sílaba tônica pode ser detectada, por exemplo, através de uma descida (F) seguida de uma subida (R) da onda fundamental. Caso a tonicidade da sílaba não seja a usual de determinada palavra, diz-se que houve um *acento léxico*. Isto ocorre normalmente por variações de dialeto ou por acentos enfáticos indicando emoções.

As **frases prosódicas** são unidades de fala limitadas por palavras que possuem variações F, R ou CR da onda fundamental, ou ainda possui as sílabas terminais estendidas no tempo. Kompe define dois tipos de delimitação de frases prosódicas: *limites por cláusulas prosódicas* (etiqueta B3) e *limites por constituintes prosódicos* (B2) (KOMPE, 1997). A limitação por cláusulas prosódicas é dada por entonação e duração, além de pausas. Já os limites por constituintes prosódicos são marcados por pequenos movimentos de entonação e nunca possuem pausas.

O conceito de **ritmo** difere com a linguagem. Em linguagens acentuadas (*stress timed*), ritmo refere-se à contagem do número de sílabas acentuadas em relação ao tempo. Já em linguagens silábicas (*syllable timed*), como o português, o ritmo depende do número de sílabas por período de tempo. A análise do ritmo pode ser importante para verificação da estrutura do discurso e para a análise semântica das palavras, uma vez que o ritmo imposto a uma pode alterar seu sentido.

A **hesitação** marca um conflito entre o planejamento da fala e sua produção. Uma pausa preenchida ou uma sílaba pronunciada de forma prolongada geralmente referem-se a uma hesitação. Quando uma hesitação é muito longa, ela é chamada de hesitação comprida (hesitation lengthening) e recebe a etiqueta B9 (KOMPE, 1997).

A etiquetagem e determinação dos elementos prosódicos constituem-se num processo minucioso, uma vez que eles ocorrem muito próximos no tempo, podem ocasionar um erro na identificação de um elemento. Também há variações que são dependentes do locutor que podem provocar erros de etiquetagem. Informações destas variações terão que acompanhar a etiquetagem para que sirvam de base na determinação de expressões não-lingüísticas envolvidas no processo de parsing da linguagem falada.

4.4.2 Identificação e classificação de características prosódicas no sinal de fala

Todo o processo de identificação de elementos prosódicos do sinal de fala é guiado pelo comportamento da onda fundamental. Desta forma, é essencial a representação adequada da envoltória dessa onda para que todo o processamento não seja comprometido.

Segundo Kompe, um bom algoritmo para determinação da freqüência fundamental é aquele baseado no fato que a harmônica de maior freqüência dividida pela fundamental resulta num valor inteiro (KOMPE, 1997). Assim, para cada quadro analisado no sinal de fala são calculados os pontos médios do espectro na tentativa de ser obtidos os pontos da fundamental. A partir dos pontos gerados é realizado o cálculo da envoltória da onda estimada. Esta envoltória da onda fundamental deverá ter um comportamento que reflete a prosódia do sinal de fala.

Com base no comportamento da onda fundamental, pode ser inferenciada a inflexão de uma sentença, ou seja, é possível identificar se uma frase pronunciada é uma afirmação, questão ou possui uma pausa (como uma indicada por vírgula ou uma hesitação). A afirmação é indicada por uma subida da onda (R); questão é marcada por uma descida (F) e a pausa é uma continuação seguida de subida (CR) (KOMPE, 1997). Este tipo de dado é essencial em diálogos, e importante para determinação de pontuação de textos. A identificação da inflexão da frase pode ser feita automaticamente pelo treinamento de uma rede neural.

Já a acentuação e os limites da frase prosódica possui uma complexidade maior para determinação de seus elementos, necessitando informações de duração e energia, além da onda fundamental. O intervalo de tempo que irá guiar o processamento pode depender de vários fatores: número de sílabas; núcleo da sílabas; palavras; limites das frases prosódicas (KOMPE, 1997). A partir da determinação de um destes parâmetros, será possível retirar do sinal as características desejadas.

As características podem ser sintetizadas em: duração, onda fundamental, energia, pausa e léxico. A duração refere-se à velocidade de fala e duração média do fonema em dado intervalo. Características da onda fundamental são a média, a mediana, seus valores mínimo e máximo, seu deslocamento no tempo, posição no tempo relativo à sílaba/palavra/frase fonética, entre outros. Da energia pode-se obter as mesmas propriedades da onda fundamental, com exceção do deslocamento em função do tempo. O tamanho da pausa pode ser determinado em milissegundos antes ou depois de uma sílaba ou palavra. Características léxicas podem ser obtidas através da classificação de um fonema do núcleo silábico, identificação da sílaba tônica ou determinação da sílaba final de uma palavra (KOMPE, 1997).

4.4.2.1 Modelagem de Entonação

Os elementos compostos da prosódia, como visto, são composições dos elementos básicos. Para sua identificação, é necessária a modelagem estatística para previsão da ocorrência dos elementos compostos. Helen Wright (hoje Helen Wright Hastie) definiu isto em sua tese com sendo uma modelagem de entonação (WRIGHT, 1999; HASTIE; POESIO; ISARD, 2002). Este tipo de modelagem serve para prever a ocorrência de indicações de surpresa, dúvida, formalidade, entre outras características que podem ser utilizada na identificação de contextos de diálogo em jogos.

Os estudos de Wright utilizaram três abordagens estatísicas para modelagem de entonação: Árvores de Classificação e Regressão (CARTs - Classification and Regression Trees), Modelos de Markov e Redes Neurais Artificiais. Segundo Wright, a primeira técnica obteve desempenho superior em seus testes de identificação de contextos de diálogo.

As CARTs são árvores binárias utilizadas para classificação dos movimentos dos jogos, dadas as características da entonação. Seu método de classificação é baseado na freqüência de ocorrência de determinada característica prosódica.

A modelagem por Markov segue os mesmos princípios que vêm sendo discutidos no decorrer deste trabalho. Nesta abordagem em particular, são modeladas combinações das características prosódicas, que são treinadas para identificar o tipo de movimento do jogo. Quanto às redes neurais, Wright utilizou-se do método backpropagation para treinamento do tipo de movimento, dadas características prosódicas de entrada. Para mais detalhes sobre sistemas markovianos e neurais, veja o capítulo 5.

Apesar das CARTs terem obtido melhor desempenho neste sistema, Wright salienta que foi apenas um pequeno ganho, o que não permite descartar as demais abordagens em outros contextos de aplicação.

4.4.3 Uso da prosódia na análise da linguagem falada

Apesar da prosódia não ser um dado explícito da linguagem, ela pode ser usada como um guia para o parsing da linguagem falada. A resolução dos problemas de ambigüidade apresentados anteriormente neste trabalho têm sua resolução apoiada pelo uso das marcas prosódicas contidas na fala.

Como visto inicialmente na seção 2.4, os sistemas de reconhecimento da fala incorporam, além do reconhecimento dos fonemas/sílabas, também a identificação das características prosódicas. Desta forma são gerados grafos de probabilidades, mas acrescidos de etiquetas prosódicas. Estas etiquetas auxiliarão na desambigüação nos níveis de sintaxe, semântica e pragmática.

Como os limites prosódicos nem sempre combinam com os limites sintáticos, é necessária a colocação de etiquetas sintático-prosódicas, chamadas de limites M (de Modelo de linguagem). Estes limites podem ser denominados limites ambíguos (MU), limites inexistentes (M0) ou limites principais (M3). Normalmente, dentro dos limites M são mapeados 5 subclasses de limites sintáticos: limite inexistente (S0), com partículas (S1), com frases (S2), com cláusulas (S3), com cláusulas principais ou frases livres (S4) (NÖTH et al., 2002).

Com relação a frases na forma de questão também pode haver disparidade prosódico-sintática. Desta forma, questões prosódicas são indicadas por PQ e questões sintáticas são indicadas por SQ (NÖTH et al., 2002). Isto é necessário, uma vez que nem sempre uma descida na onda fundamental (etiqueta prosódica F) indica uma questão na sua forma sintática.

Não foi encontrada na literatura so processamento da linguagem falada a existência de etiquetas semânticas, mas há etiquetas pragmáticas, utilizadas para indicar limites de contextos de diálogos, também chamados de *atos de diálogo*. Estas etiquetas dependem do sistema, mas em geral D0 indica a ausência de limite e D*i* é o diálogo do contexto *i* (HASTIE; POESIO; ISARD, 2002; NÖTH et al., 2002).

Para a inserção das etiquetas prosódicas, há diferentes formas de processamento que podem ou não estarem relacionadas. Há técnicas de análise de atos de diálogo que não realiza o parsing prosódico-sintático, por exemplo. Apesar disso, a compreensão da linguagem falada envolve todas as etapas de parsing, que serão analisadas a seguir.

4.4.3.1 Prosódia na análise sintática

Uma vez que um sistema que use prosódia se utiliza deste conhecimento para melhor realizar o reconhecimento das palavras, os grafos gerados já possuem etiquetas prosódicas que guiarão o resto do processo. Ou seja, uma vez reconhecidas as hipóteses de palavras, é necessário um caminhamento no grafo de hipóteses para que seja encontrada a melhor seqüência de palavras.

A varredura do grafo de hipóteses pode ser feita por um algoritmo como A* ou de Viterbi. Kompe indica que A* possui desempenho superior neste tipo de processamento (KOMPE, 1997).

A sintaxe utiliza-se da prosódia basicamente para duas operações: verificação da acentuação e otimização da busca da melhor seqüência de palavras (KOMPE, 1997). A acentuação simplesmente verifica as etiquetas prosódicas a ela referentes. Já a seqüência de palavras terá de ser devidamente pontuada, estabelecendo as frases sintáticas. Para tanto, serão utilizadas as etiquetas de frases prosódicas e pausas anteriormente identificadas.

A partir das características prosódicas, também podem ser identificados os tipos de frases e quebras da seqüência de fala. Por exemplo, seja a frase falada: *vamos nos encontrar na ter... na quarta*. Obviamente que a hesitação encontrada entre *ter(ça)* e *na* será um fator determinante para uma adequada correção sintática: *vamos nos encontrar na quarta*.

Cada sistema desenvolvido pode utilizar as características prosódicas que forem úteis para sua análise sintática. Alguns sistemas que utilizam a prosódia para auxílio na análise sintática foram citados na seção 2.4.

4.4.3.2 Prosódia na análise semântica

Para a resolução de ambigüidades semânticas, utiliza-se da etiquetagem prosódico-sintática. Para dada seqüência de palavras, é verificada a acentuação do conjunto. Com base em regras, são avaliadas as melhores hipóteses de frases, segundo sua fluência.

Segundo (KOMPE, 1997), a prosódia atua na análise semântica no que se refere ao *foco* de uma fala. Na linguagem falada, o foco ou acento focal indica a importância de determinada palavra na sentença. Por exemplo, na frase *vamos lá* ***hoje***, a ênfase dada à palavra *hoje* define urgência, tendo, portanto, um significado diferente da mesma frase sem a ênfase.

Mesmo numa mesma frase, sua pronúncia pode ter diferentes focos, ou seja, o foco pode ser deslocado. No início deste capítulo foi dado um exemplo de deslocamento de foco, com a frase *o professor chegou*. O foco pode estar em *professor*, em *chegou* ou simplesmente não existir. Cada uma destas hipóteses gera um significado diferente para uma mesma frase.

Outra informação semântica, já discutida anteriormente, é a inflexão de uma frase.

Esta pode ter o sentido de uma questão ou afirmação, o que muda completamente o sentido da frase: *agora posso ir?* é muito diferente de *agora posso ir*. Informações como esta são utilizadas também na definição de atos de diálogo na análise pragmática.

A análise semântica resulta em mais anotações no grafo com as hipóteses de frases, o qual, por sua vez, será analisado a nível pragmático. O processamento realizado pela desambigüação semântica será muito relevante para a escolha do contexto de diálogo.

4.4.3.3 Prosódia na análise pragmática

A prosódia auxilia na análise pragmática através da determinação de contextos (atos) de diálogo. Características prosódicas como duração, pausa, onda fundamental, energia e gênero são classificadas em árvores de decisão probabilísticas. Alguns sistemas que utilizam a prosódia para análise pragmática foram mencionados na seção 2.4.

4.4.4 Extração de características prosódicas através de ondeletas

O objetivo da extração das características prosódicas nesta tese é a obtenção de informações para análise semântica. Ao contrário da prosódia lingüística, que identifica pausas e hesitações na sintaxe, está sendo abordada a prosódia afetiva, que identifica a carga emocional e não verbal contida na fala (BOSTANOV; KOTCHOUBEY, 2003). Nesta tese não se pretende, num primeiro momento, a identificação explícita dos elementos prosódicos, mas sim do agrupamento daquelas palavras que possuem similaridade entre estes elementos.

A análise do pitch pode ser realizada por meio de ondeletas e daí realizar-se a obtenção de características prosódicas. Segundo (KADAMBE; BOUDREAUX-BARTELS, 1990, 1992), o pitch pode ser estimado localizando-se o instante que é feito o fechamento da glote. O tempo medido entre os intervalos de fechamento da glote é, portanto, um evento de detecção do pitch. Kadambe observa que o instante de fechamento da glote (GCI - *Glottal Closure Instant*) pode ser calculado como sendo o ponto de máximo de uma matriz de autocovariância do sinal. No entanto, ele observa que este método falha para sinais não-estacionários como a fala.

Ainda segundo (KADAMBE; BOUDREAUX-BARTELS, 1990, 1992), métodos clássicos de estimação do pitch necessitam extrair os coeficientes cepstrais e buscar a diferença de magnitude média da autocorrelação destes coeficientes. Os problemas apontados por Kadambe nestes métodos são o período do pitch com relação ao tamanho do segmento (fixo) e as diferenças entre locutores com pitch alto e baixo.

Como alternativa aos métodos clássicos, Kadambe propõe o uso de uma ondeleta diádica, a qual realiza o processamento como o descrito na análise em multiresolução, seção 4.3.2. As vantagens deste método seria a tolerância a deslocamentos do sinal e permitir a identificação do máximo local em torno de pontos de descontinuidade, o que representa a variação do fluxo de ar na fala (KADAMBE; BOUDREAUX-BARTELS, 1992). Em estudos recentes, Tran demonstrou que a fase da transformada discreta de ondeletas é um sinal periódico e possui o mesmo período do pitch (TRAN; HA; DISSANAYAKE, 2004).

O processamento proposto por Kadambe identifica os pontos de máximo em cada escala que excedem a 80% do máximo da escala anterior. Caso os pontos coincidam em duas escalas, ou seja, do sinal atual e sua derivada, eles são considerados um GCI. Segundo Kadambe, bastam 3 escalas para uma adequada estimação do pitch (KADAMBE; BOUDREAUX-BARTELS, 1992).

A partir do estudo de Kadambe muitos outros derivaram-se, em busca de métodos mais otimizados. Este é o caso do trabalho de Bruin, que afirmou ser possível a obtenção

de dois grupos de características prosódicas: um relativo ao tom, que indica a variação das sílabas, palavras ou frases, e outro referente à entonação de frases (BRUIN; PREEZ, 1993). Ele identificou períodos de pitch e a partir destes a língua falada por locutores.

A otimização da identificação do pitch resultou em outros estudos, como o de (EVANGELISTA, 1993), que desenvolveu ondeletas específicas para o problema de sincronização do pitch de diferentes escalas. Segundo Evangelista, seu método permite uma localização mais precisa do pitch, diferenciando-o das harmônicas do sinal.

Outra abordagem é feita por (OBAIDAT et al., 1999), que propõe a comparação do ponto de máximo da potência espectral com o os máximos encontrados pela transformada ondeletas. Os valores coincidentes resultantes deste processo formarão os GCIs para estimação do pitch.

Segundo Chen, ocorre um efeito de *aliasing* (veja seção 4.1) durante a decimação no processo de filtragem na análise em multiresolução (CHEN; WANG, 2001). Para resolver este problema, utiliza-se um filtro anti-aliasing a cada decimação, restaurando as diferenças espectrais e permitindo uma melhor identificação do pitch.

Para melhorar a detecção do pitch em situações de ruído, Gavat propôs o uso da transformada de Hilbert para retirar os falsos pontos de máximo (GAVAT; ZIRRA; SABAC, 2002). O ruído promove pontos de máximo falsos (que não são GCIs) nos coeficientes das ondeletas, mas em regiões de silêncio ou não vocálicas. Segundo Gavat, a transformada de Hilbert ameniza este efeito.

4.5 A codificação do sinal de fala como ferramenta de representação para a linguagem e a prosódia

Neste capítulo foi apresentada uma visão onde uma transformada matemática, como ondeletas, pode ser usada para extração de características da linguagem falada e da forma como esta linguagem é expressada. A propriedade das ondeletas de poder ser usada como um filtro passa-banda permite a codificação da faixa de freqüência de áudio que contém a voz.

De posse desta ferramenta é possível a obtenção das características do sinal em função do tempo. Observa-se que aqui não foram tratados aspectos sobre a segmentação da fala, para melhor análise do tempo. Apesar da segmentação da voz poder ser tratada pela análise da prosódia, não é alvo de estudo na presente Tese.

Como analisado, as características do sinal extraídas por meio da transformada ondeletas podem ser utilizadas tanto para representação da linguagem como para estimação da onda fundamental (F0). Com dados da F0, é possível a análise da prosódia. Em suma, a transformada ondeletas permite tanto a representação da linguagem falada como da prosódia. No próximo capítulo serão descritas técnicas de como os dados obtidos a partir do sinal de voz podem ser processados para análise da linguagem falada.

5 ANÁLISE DA LINGUAGEM FALADA

5.1 Processamento de Linguagem Natural

Normalmente, um sistema de processamento de linguagem natural é abordado do ponto de vista da análise do conhecimento morfológico, sintático, semântico e pragmático. A análise morfológica estuda a construção das palavras, com seus radicais e afixos, que correspondem a partes estáticas e variantes das palavras, como as inflexões verbais. A análise sintática diz respeito ao estudo das relações formais entre palavras (JURAFSKY; MARTIN, 2000). A análise semântica é um processo de mapeamento de sentenças de uma linguagem visando a representação de seu significado (ABRAHÃO; LIMA, 1996). A pragmática diz respeito ao processamento da forma que a linguagem é utilizada para comunicar, como os significados obtidos na análise semântica agem sobre as pessoas e seu contexto.

Para a implementação de analisadores sintáticos é necessária uma prévia classificação das palavras para posterior verificação da sua posição na frase. No português, por exemplo, temos 10 classes básicas de palavras: substantivo, verbo, adjetivo, artigo, numeral, pronome, advérbio, preposição, conjunção e interjeição (CUNHA, 1982). Cada uma dessas palavras possuem combinações coerentes com a sintaxe definida pela linguagem. Não é coerente, por exemplo, a colocação de uma conjunção antes de uma preposição. A verificação disto é tarefa da análise sintática.

Para a análise semântica são definidos modelos de léxico onde são associados significados a palavras ou expressões. Na implementação de analisadores semânticos são utilizadas diversas técnicas, como redes semânticas, gramáticas semânticas e modelos conexionistas (em geral, redes neurais).

A análise pragmática geralmente é implementada através de quadros (*frames*), casos (*cases*) e roteiros (*scripts*), associando os significados obtidos a um determinado contexto. Normalmente casos de ambigüidade da linguagem são resolvidos neste nível, uma vez que o contexto permite a diferenciação necessária.

Apesar das diversas etapas de processamento da linguagem, há pesquisadores que acreditam ser muito tênue a separação entre reconhecimento sintático, semântico e pragmático. Neal e Shapiro afirmam que alguns aspectos da sintaxe não deveriam ser separados da semântica, uma vez que a mudança de uma palavra que está na frente de outra pode alterar-lhe o sentido. Por exemplo a palavra *areia* na frase *Eu usei areia*, ao colocar um artigo na frente, muda seu significado: *Eu usei a areia*. A primeira frase indica uma areia qualquer, mas a segunda especifica uma determinada areia (NEAL; SHAPIRO, 1987).

Por outro lado, a semântica também tem íntima relação com a pragmática, uma vez que há palavras que somente encontram seu significado no contexto, como por exemplo a palavra *banco* na frase *Eu tropecei no banco*. Dependendo do contexto, *banco* pode

Tabela 5.1: Tabela de sufixos.

sufixo	**substantivo (radical)**	**palavra derivada**
-zinho	boné	bonezinho
	botão	botãozinho
-zinha	árvore	árvorezinha
	flor	florzinha

indicar o móvel ou a agência bancária. Este tipo de ambigüidade só pode ser resolvida a nível de análise pragmática.

No decorrer deste capítulo serão abordadas todas as fases do processamento da linguagem natural, desde a análise morfológica até a análise de discurso.

5.1.1 Análise morfológica

Este tipo de análise é necessário para que o tamanho do dicionário não fique muito extenso, uma vez que é mais simples o armazenamento do radical da palavra e seus afixos. Estes são os componentes que formam uma palavra juntamente com o radical, como prefixos e sufixos. Um exemplo de prefixo é *des* na palavra *desesperança* e de sufixo é *mos* na palavra *calamos*.

O tratamento computacional deste tipo de análise é relativamente simples. Baseia-se em regras que analisam as palavras e as classificam segundo tabelas de afixos. Por exemplo, a entrada *zinho* de uma tabela de sufixos está associada a um diminutivo de um substantivo, portanto, a palavra *bonezinho* é o diminutivo da palavra *boné*, que é seu radical. Desta forma, são reconhecidas as palavras que não estão na sua forma padrão, já adequando-as para a fase posterior de análise sintática. A tabela 5.1 apresenta o uso de dois sufixos.

Jurafsky e Martin chamam o sistema que realiza a passagem da palavra escrita para a forma classificada de *Conversor de Estados Finitos* (Finit-State Transducer - FST). Este tipo de sistema é o que passa uma palavra de sua forma como é normalmente escrita para uma forma *etiquetada*, ou seja, com identificação de seu radical e afixo (JURAFSKY; MARTIN, 2000). Na seção 5.1.2.1 será apresentado o nível de etiquetagem sintática, complementar à morfológica.

Um exemplo de analisador morfológico para o português é o sistema Palavroso, que realiza a identificação do afixo e pré-classifica as palavras em quatro grupos: verbos, nomes e adjetivos, classes fechadas e advérbios (WITTMANN; RIBEIRO, 1998). A implementação desse sistema tem por base o que foi anteriormente descrito, mas possui ainda outro módulo de dicionário que verifica a existência da palavra construída através das regras morfológicas.

5.1.2 Análise sintática

No contexto do processamento da linguagem, as gramáticas utilizadas na análise sintática têm sido chamada de Modelos de Linguagem (Language Models - LM). Esta definição está associada ao amplo universo de frases possíveis de serem modeladas para representação de determinado domínio de análise.

Tabela 5.3: Etiquetagem sintática.

etiqueta	**descrição**	**palavra**
PPE	**P**ronome **PE**ssoal	eu
VP	**V**erbo no **P**assado	tropecei
PAF	**P**reposição+**A**rtigo **F**eminino	na
SSF	**S**ubstantivo **S**ingular **F**eminino	pedra

No seguimento, serão abordadas algumas técnicas utilizadas na identificação e etiquetagem dos elementos sintáticos e na organização da construção sintática.

5.1.2.1 Etiquetagem

O primeiro processamento que é efetuado na análise sintática é a identificação das classes das palavras (também conhecidas como classes morfológicas, etiquetas lexicais ou partes de fala). Para proceder esta classificação, são implementados parsers que identificam nas frases as classes das palavras que as compõem. Esta classificação de palavras também é conhecida como etiquetagem (tagging).

Por exemplo, a frase *Eu tropecei na pedra*, poderia ser etiquetada da seguinte forma: Eu/PPE tropecei/VP na/PAF pedra/SSF. Esta representação segue a tabela 5.3, onde são apresentadas as etiquetas, sua descrição e a palavra correspondente na frase apresentada.

Basicamente, a etiquetagem dos atuais parsers sintáticos que vem sendo implementados constituem-se de três tipos: os que usam regras, os que utilizam estatísticas e os híbridos, que utilizam as duas anteriores (PACHECO; DILLINGER; CARVALHO, 1996; JURAFSKY; MARTIN, 2000). A primeira, que utiliza regras, também é chamada de *lexicalista*, uma vez que ela se preocupa em seguir regras de dicionário e de gramáticas para verificação da consistência da linguagem.

A segunda abordagem chama-se *probabilística*, uma vez que utiliza cadeias de Markov (veja seção 5.3) para descobrir qual é a seqüência mais provável de palavras. Dentro deste contexto, a técnica que tem sido mais utilizada em parsers sintáticos probabilísticos é a de *n-gramas*, que consiste em estabelecer uma estatística a cada *n* palavras. Normalmente é implementado n=3, sendo que são analisadas 2 palavras para a previsão da terceira. Desta forma, é estabelecida uma estatística de construção de frases, permitindo a descoberta de qual é a palavra de combinação mais provável, permitindo a verificação e correção da construção frasal.

A última abordagem é chamada de *funcionalista* e reúne tanto características referentes ao uso de regras, quanto ao uso de probabilidades. Uma das técnicas mais utilizadas é conhecida como Transformation-Based Learning (TBL), ou Aprendizado Baseado em Transformação, que faz a indução de regras a partir de exemplos de palavras apresentadas. Baseando-se em regras básicas, procura-se inferir a categoria de construção das frases dadas para aprendizado. Sobre este aprendizado, são construídas estatísticas que podem dar origem a novas regras de construção de frases não previstas no modelo básico.

A etiquetagem das palavras, contudo, não basta para a análise sintática. Por vezes ocorrem situações ambíguas onde é necessário recorrer-se a mais um nível, uma vez que, dependendo onde se encontra a palavra na frase, ela pode ter a função de advérbio ou

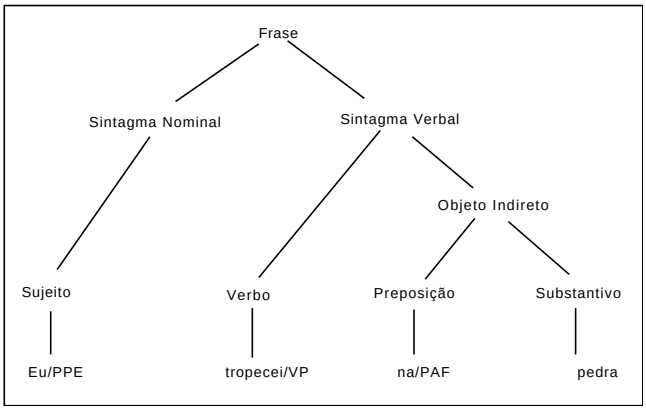

Figura 5.1: Árvore de parser.

de substantivo (ex.: *branco*), de verbo ou conjunção (ex.: *como*), e assim por diante. O próximo passo, portanto, é a análise da construção gramatical da frase para a definição da categoria sintática.

5.1.2.2 Construção Gramatical

Após a etiquetagem, passa-se à verificação da construção gramatical da frase. Para tanto, as linguagem é modelada por gramáticas livres de contexto, que divide as frases em árvores de sintagmas nominal e verbal e, a partir daí, verifica que classe enquadra-se em cada sintagma (veja figura 5.1). Neste sentido, há duas correntes de desenvolvimento: a construção de árvores de parser e de parsers probabilísticos (JURAFSKY; MARTIN, 2000).

Árvores de Parser

Árvores de Parser utilizam técnicas de busca em árvore para determinar a correção da construção frasal, realizando a comparação entre a frase e a estrutura da árvore. Uma técnica bastante difundida que busca resolver os problemas de busca em árvore é o algoritmo de Earley, que baseia-se na transição de estados para estabelecer qual é a classe mais provável de se seguirá à classe atual (SIKKEL, 1997; JURAFSKY; MARTIN, 2000). Este algoritmo possui 3 operadores: *predictor*, *scanner* e *completor*. O primeiro estabelece as transições de estado possíveis, o segundo realiza a comparação entre as transições possíveis e a palavra em análise, e o terceiro módulo realiza o registro final da classe na frase etiquetada. O maior problema deste tipo de parser é a incapacidade de resolver casos de ambigüidade.

Algoritmo de Earley

O algoritmo de Earley realiza uma busca top-down numa gramática estruturada em árvore. Ele começa com a varredura de um vetor chamado *mapa* (chart), contendo tantos elementos quantos forem as palavras da frase em análise. Cada uma destas entradas do vetor contém uma lista de estados gerados pelo parsing até aquela palavra. A lista de estados gerados é composta de uma subárvore que representa uma regra gramatical, de informação sobre o preenchimento da subárvore (regras satisfeitas) e da posição na subárvore de acordo com a palavra do mapa (vetor).

Neste contexto, o operador *predictor* é que cria a subárvore a partir da regra grama-

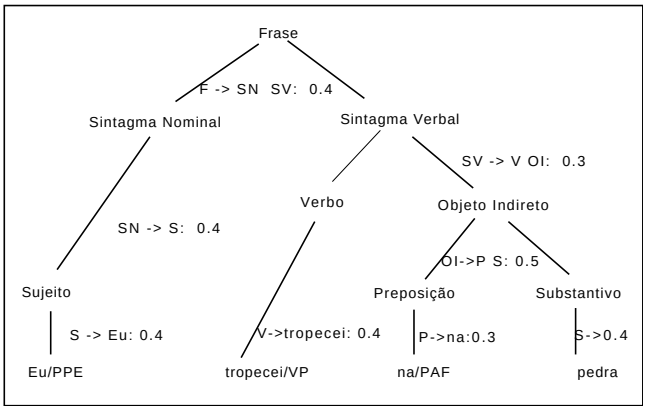

Figura 5.2: Árvore de parser com indicações de probabilidade.

tical. Ele aloca as folhas na árvore para estados não-terminais da gramática. O operador *scanner* verifica qual das ramificações geradas é compatível com a palavra em análise, gera uma nova folha e marca no *mapa* o atual estado da busca. Por fim, o operador *completer* é necessário quando uma regra terminou de ser avaliada e deve-se realizar o retrocesso (backtracking) para concluir a avaliação das demais ramificações da árvore.

Parsing Probabilístico

O parsing probabilístico utiliza uma gramática cujas entradas possuem ponderações estatísticas (pesos). Estas ponderações são colocadas com base na observação da probabilidade de ocorrência das regras gramaticais no *corpus* (base de palavras). Através destes pesos é possível a resolução de ambigüidades, sendo indicadas as frases de maior probabilidade.

Assim, a probabilidade de uma regra R não terminal ($a \rightarrow b \mid a$) ocorrer numa frase F é a divisão do número de ocorrências de R na avaliação de F pelo número de repetições do símbolo avaliado (a):

$$P(R) = P(a \rightarrow b \mid a) = \frac{cont(a \rightarrow b)}{cont(a)}$$

E a probabilidade de uma árvore de parser A para o cálculo da melhor F é o produto de todas as regras envolvidas na construção de A:

$$P(A) = \prod_{n \in A} P(Rn)$$

A partir destas equações é possível saber se as regras usadas para a construção de uma árvore é mais adequada que outra. Veja um exemplo de árvore com probabilidades na figura 5.2. Isto resolve grande parte dos problemas de ambigüidade nos casos das palavras que podem pertencer a mais de uma classe, como visto na seção 5.1.2.1.

Um dos algoritmos empregados para a execução deste tipo de parsing é o Earley probabilístico, que segue o mesmo algoritmo analisado na seção anterior, acrescido dos valores de probabilidades das regras, o que permite a desambigüação.

O maior problema do parser probabilístico é a colocação de pesos iguais a construções frasais diferentes, embora com as mesmas palavras *(ex.: eu tropecei na pedra* poderia ter

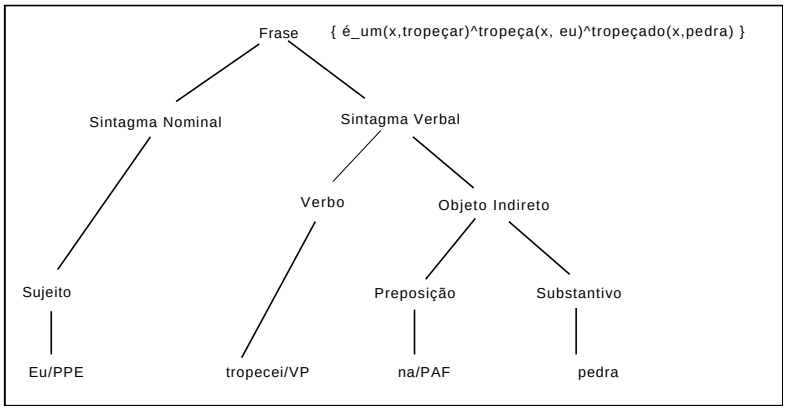

Figura 5.3: Árvore de parser com anexo semântico.

igual peso de *tropecei na pedra eu*). A solução para este caso está na utilização de indicadores lexicais (*lexical heads*), que permite ao algoritmo de busca identificar diferentes ramificações dos sintagmas.

Os indicadores lexicais são palavras que ficam junto das raízes de uma determinada subárvore de parsing. Isso serve para indicar a palavra principal de uma parte da frase (o sujeito num sintagma nominal, por exemplo) e alterar o peso de construções frasais que poderiam ter pesos iguais.

5.1.3 Análise Semântica

Apesar do extenso processamento realizado nas análises morforlógica e sintática, apenas com estas não é possível distinguir certas categorias de palavras e muito menos precisar o objetivo da frase. Para tanto, são acrescentadas nas árvores de parser os chamados *anexos semânticos* (semantic attachments) (JURAFSKY; MARTIN, 2000). Este tipo de análise semântica tradicional pode ser construído ainda em tempo de análise sintática, à medida que a árvore de parser vai sendo completada (veja figura 5.3).

O anexo semântico apresentado na árvore da figura 5.3 é uma composição de uma rede semântica onde é definido o verbo *tropeçar*. A definição diz que este verbo necessita de um agente (aquele que tropeça) e um paciente (algo em que o agente tropeça).

Por outro lado, há outras formas de análise semântica: gramáticas semânticas (*semantic grammar*), gramáticas baseadas em caso (*case-based grammars*) e outros métodos para casos específicos (CARBONEL; HAYES, 1987).

5.1.3.1 Gramáticas Semânticas

As gramáticas tradicionais são utilizadas apenas para definir regras de sintaxe. Por outro lado, podem ser definidas regras associadas a frases-padrão, onde variam as palavras envolvidas. É uma forma viável quando existe uma hierarquia de diferentes contextos e necessita-se diferenciá-los através de padrões. Infelizmente, para cada frase analisada, é necessária uma regra distinta, dificultando sua utilização em diferentes domínios do conhecimento (CARBONEL; HAYES, 1987; PACHECO; DILLINGER; CARVALHO, 1996; MINKER, 1998; JURAFSKY; MARTIN, 2000).

A construção de gramáticas semânticas baseia-se na colocação de variáveis que serão preenchidas de acordo com a afirmação ou questionamento do usuário ao sistema. Um exemplo de regras de uma gramática semântica poderia ser:

```
QUESTÃO -> quem tropeçou ONDE
RESP_QUEST -> QUEM tropeçou ONDE
QUEM -> eu
ONDE -> PREPOSIÇÃO LUGAR
PREPOSIÇÃO -> na
LUGAR -> pedra
```

Caso fosse questionado *Quem tropeçou **na pedra**?*, a resposta seria ***eu** tropeçou na pedra*, pela associação da expressão *na pedra* com a variável ONDE e o pronome *eu* com a variável QUEM.

5.1.3.2 Gramáticas Baseadas em Casos

Esta forma de análise usa gramáticas semânticas não-terminais para formar padrões. Caso uma dada frase *encaixe* na construção padrão, ela poderá ser reconhecida dentro de um contexto. Este contexto em geral é estruturado na forma de quadros (*frames*) e casos (*cases*) (CARBONEL; HAYES, 1987; BARKER, 1998; MINKER, 1998).

São construídas gramáticas semânticas para identificar os elementos da frase e indicar o caso (ou quadro) que melhor se aplica a estes elementos. Um exemplo de gramática, para uma aplicação de leitura de correio eletrônico, pode ser o seguinte:

```
<mensagem> -> <?artigo cabeça-mensagem *caso-mensagem>
<caso-mensagem> -> <%de Pessoa>
<caso-mensagem> -> <%sobre Assunto>
<caso-mensagem> -> <%desde Data>
```

Onde *?* indica um item opcional, * um item passível de repetição e % significa que qualquer palavra da mesma classe daquela indicada a seguir pode ser utilizada; *artigo* indica que pode haver um artigo antecedendo a cabeça da frase; *cabeça-mensagem* é a palavra-chave que indica ser uma mensagem do correio eletrônicos; *caso-mensagem* é a referência para o caso que irá tratar daquela parte da frase; *Pessoa*, *Assunto* e *Data* são variáveis que recebem os dados sobre a mensagem.

Uma frase que poderia se encaixar neste modelo seria *as mensagens de João sobre modelagem desde segunda-feir*a. Neste caso, as palavras *João*, *modelagem* e *segunda-feira* servirão de guia para o caso a ser utilizado na resposta do sistema.

A construção de casos para análise semântica antecipa um tipo de processamento geralmente utilizado apenas na análise pragmática, usado para descrever contextos específicos. Na implementação de casos são utilizados *quadros*, que são estruturas utilizadas para caracterizar os elementos semânticos que compõem a frase em análise. *Casos* são, portanto, estruturas contendo quadros utilizados em dado contexto que identificam problemas e ainda outros quadros que identificam soluções para estes problemas. Estas estruturas podem ser usadas na análise semântica para resolver ambigüidades não esclarecidas no processo convencional, baseado em regras.

Um caso para o exemplo citado anteriormente poderia ser o seguinte:

```
[
 Nome: Mensagem
 Esquema:
  [
   Emissor: &Pessoa
```

```
   Assunto: &Assunto
   Data: &Data
  ]
 Sintaxe:
  [
   Tipo: Frase-nominal
   Cabeça: (mensagem mensagens mail carta)
   Caso:
    (
     <%de Pessoa>
     <%sobre Assunto>
     <%desde Data>
    )
  ]
]
```

Por outro lado, há quem argumente que a análise pragmática deva ser feita junto com a análise semântica através do uso de casos . A justificativa para isso está na abordagem do estudo das relações existentes na frase de um caso. Dependendo da relação, define-se a função da frase dentro do caso, obtendo-se, ao mesmo tempo, o aspecto semântico (significado) e pragmático (contexto).

5.1.4 Análise Pragmática

Este tipo de análise foge à estrutura de apenas uma frase. Ela busca nas demais frases a compreensão do contexto que falta à frase em análise. Em geral, não há estruturas pré-definidas que atendam a uma representação adequada de problemas de referências pronomiais (por exemplo os pronomes *la, seu* e *o* no contexto: *João pegou a rosa. Ao pegá-la, seu espinho o espetou*), coerência textual e análise de discurso. Por outro lado, as estruturas mais utilizadas na análise pragmática são os casos, conforme descritos na seção 5.1.3.2, que por sua característica já permitem a identificação de contexto.

Para se ter uma noção dos algoritmos utilizados na análise pragmática, pode-se pegar um conhecido algoritmo de resolução de referências pronomiais, proposto por Lappin e Leass em 1994 (JURAFSKY; MARTIN, 2000). Neste algoritmo, são atribuídos pesos aos pronomes encontrados em frases. Esses pesos são chamados de *valores de saliência* e são atribuídos aos sujeitos encontrados na mesma frase ou nas frases precedentes. Quanto mais próximo do pronome estiver o sujeito encontrado, maior será o valor a ele atribuído como referência ao pronome.

Após serem atribuídos pesos aos sujeitos de referência encontrados, retiram-se aqueles que não combinam em gênero e número, e são levadas em conta outras características como o tipo de pronome, se a frase tem objeto indireto ou advérbio, entre outros fatores. Cada uma dessas características podem influir no cálculo do peso da referência, de forma a que o sujeito com maior peso será o selecionado para substituição do pronome. Assim, num contexto como o exemplo citado anteriormente, o pronome *la*, *seu* e *o* terá como sujeito mais próximo *a rosa*, mas o último pronome (*o*) não combina com o gênero, sendo, então, atribuído ao próximo sujeito, *João*.

Há ainda muitos outros algoritmos citados na literatura acerca de resolução de referências pronomiais, porém a maioria é dependente de uma estrutura sintática previamente determinada ou pouco flexível a diferentes construções frasais. Por esta dificuldade de

ser identificada a construção frasal é que tem sido cada vez mais usadas estruturas de preenchimento como *casos* para análise pragmática.

5.2 Processamento da Linguagem Natural Falada

Todos os pesquisadores que desenvolvem sistemas para a linguagem falada concordam que sua análise é mais complexa que a linguagem escrita. Por esta razão, afirma-se que a análise da linguagem escrita é um subconjunto da análise da linguagem falada, uma vez que esta pode ser falada seguindo as regras gramaticais correntes na língua.

Normalmente, os parsers de linguagem falada são construídos a partir de regras gramaticais básicas que são otimizadas pelo sistema através de esquemas probabilísticos, de forma a adaptar a máquina de estados finitos da gramática original. Após a adaptação, são realizados diversos procedimentos de identificação de elementos genéricos, de forma a forçar a identificação de uma construção frasal ou contexto semântico, o que geralmente é realizado através de estruturas de quadros (frames).

As características extraídas do sinal de fala podem ser classificadas através de modelos probabilísticos derivados das Redes de Bayes. As redes bayesianas são conhecidas por sua capacidade de inferência estatística, nas quais são estabelecidas relações de probabilidades entre os nodos da rede, que constituem estados de um autômato finito.

Redes bayesianas podem funcionar como classificadores estatísticos de funções paramétricas. O objetivo é ajustar os parâmetros destas funções de forma a representar os vetores de padrões de características do sinal de fala (KOMPE, 1997). O classificador realiza, com base na probabilidade de um vetor V pertencer a uma classe C, a minimização do erro de representação dos parâmetros.

Dentro do processo de reconhecimento de fala, dois tipos de classificadores bayesianos são utilizados: Modelos Ocultos de Markov (HMMs - *Hidden Markov Models*) e Redes Neurais Artificiais (ANNs - *Artificial Neural Networks*). Redes neurais são excelentes para realizar a identificação dos padrões acústicos, enquanto modelos de Markov ocultos descrevem com eficácia a seqüência temporal de ocorrência dos padrões identificados pela rede neural (TEBELSKIS, 1995; MORGAN; BOULARD, 1995; KOMPE, 1997).

Pelas características descritas, os classificadores neural e markoviano geralmente são utilizados em conjunto na forma de sistemas híbridos, combinando suas potencialidades. Todavia, há modelos de redes neurais que permitem o registro da seqüência temporal, com a vantagem de não exigirem uma modelagem tão rígida quanto a abordagem markoviana. Ambos modelos de classificadores, utilizados para reconhecimento de padrões de fala, serão apresentados nas seções a seguir.

5.3 Modelos Ocultos de Markov

Autômatos finitos são comumente utilizados na Computação para representar gramáticas para a definição de linguagens. Os autômatos são representados por grafos, contendo arcos estabelecendo relações entre nodos. Caso sejam colocados pesos nos arcos, estaremos atribuindo probabilidades para estas transições, o que permite então a definição de gramáticas probabilísticas. Este tipo de representação estocástica através de grafos valorados também é conhecida como *cadeia de Markov*.

Segundo (JURAFSKY; MARTIN, 2000), estas gramáticas probabilísticas são conhecidas no reconhecimento de fala como Modelos de Linguagem (LM - *Language Model*).

Estes modelos também são conhecidos como *n-gramas*, que consiste em estabelecer uma estatística a cada *n-1* palavras. Assim, se n=2, temos um bigrama, onde a probabilidade de ocorrer a palavra atual é definida pela palavra anterior. Um bigrama também é chamado de modelo de Markov de primeira ordem, um trigrama é um modelo de segunda ordem e assim sucessivamente (JURAFSKY; MARTIN, 2000).

A probabilidade P de uma palavra p_n ocorrer dadas todas as palavras anteriores pode ser aproximada pelas N palavras anteriores (JURAFSKY; MARTIN, 2000):

$$P\left(p_n|p_1^{n-1}\right) \approx P\left(p_n|p_{n-N+1}^{n-1}\right) \tag{5.1}$$

Os modelos de N-gramas podem ser treinados pela soma das probabilidades de ocorrência (contagem C) de todos N-gramas em função da palavra p_n pela contagem das probabilidades individuais, ou seja (JURAFSKY; MARTIN, 2000):

$$P\left(p_n|p_{n-N+1}^{n-1}\right) = \frac{C\left(p_{n-N+1}^{n-1}\, p_n\right)}{C\left(p_{n-N+1}^{n-1}\right)} \tag{5.2}$$

onde é feita uma normalização de todas as N freqüências da faixa de interesse. Esta razão é chamada de *freqüência relativa* ou *estimação por máxima verossimilhança* (MLE - *Maximum Likelihood Estimation*) (JURAFSKY; MARTIN, 2000).

A modelagem por n-gramas pode ser utilizada quando a seqüência de entrada é conhecida. Mas no caso do reconhecimento de fala, as entradas são uma incógnita a serem reconhecidas. Para este tipo de situação utilizam-se modelos ocultos de Markov.

Os modelos ocultos de Markov são redes bayesianas com variáveis ocultas (BILMES, 1999). Um exemplo clássico de modelagem oculta markoviana é a estimação de um cara-ou-coroa, onde joga-se uma moeda e há 50% de chance de ser cara e 50% de ser coroa. Neste caso, as variáveis são ditas *observáveis*. Mas, caso o cara-ou-coroa esteja sendo jogado atrás de uma cortina e não seja possível saber quantas moedas estão sendo utilizadas, então a modelagem necessita de variáveis *ocultas* (RABINER; JUANG, 1993).

A modelagem do jogo cara-ou-coroa necessita da definição preliminar de quantos estados possuem o sistema: se dois (para uma moeda ou duas), três (para 3 moedas) ou mais. Para saber quais dos modelos (para 1, 2, 3 ou mais moedas) é o mais adequado e, conseqüentemente, quais estados foram utilizados, deve-se escolher qual deles adequa-se às seqüências de resultados a cada jogada. Por fim, deve-se descobrir qual é a probabilidade de cada estado ocorrer, para que seja possível a classificação de futuras seqüências. Nesta última etapa são realizados ajustes (treinamento) das variáveis de probabilidades da ocorrência de cada estado (RABINER; JUANG, 1993).

Para o caso do reconhecimento de fala, assume-se que o sinal de fala pode ser caracterizado como um processo randômico com parâmetros que podem ser estimados de forma a prever adequadamente uma seqüência de fonemas. No processo de reconhecimento, várias seqüências são avaliadas e a melhor delas (com mais alta probabilidade) será convertida numa seqüência de símbolos (caracteres) (RABINER; JUANG, 1993; JURAFSKY; MARTIN, 2000).

Um modelo oculto de Markov é formado por um conjunto *N* de estados. Como estes são ocultos, criam-se modelos baseados em dados empíricos. Como no exemplo do cara-ou-coroa existem suposições de quantas moedas existem atrás da cortina, no caso do reconhecimento de fonemas vão ser modeladas seqüências de palavras que se pretende reconhecer.

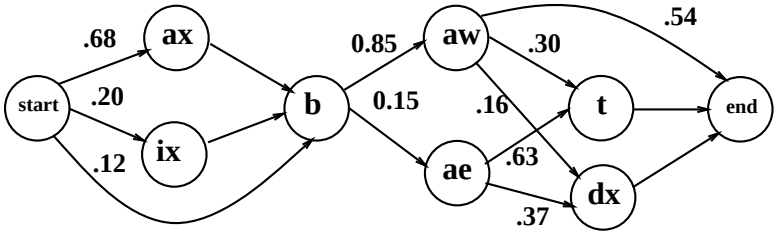

Figura 5.4: Exemplo de modelagem HMM para a palavra *about*.

Na figura 5.4 tem-se um exemplo de uma modelagem para a palavra de língua inglesa *about* (sobre), segundo (JURAFSKY; MARTIN, 2000).

Para se indicar a ordem de avaliação dos estados modelados devem ser também definidas as transições entre os estados existentes do sistema. Estas transições são organizadas numa matriz *A* de probabilidades de transição entre estados, ou seja onde e_t é o estado avaliado no tempo t. Estas transições vão indicar, por exemplo, que é mais provável de ocorrer a seqüência de letras *ba* do que *bx*.

Para se definir uma determinada seqüência de estados que ocorrerão, utiliza-se um vetor *B* de probabilidades de *M* símbolos observados, ou seja

$$B = b_j(k) = P[o_t = s_k | e_t = j], \qquad 1 \leq k \leq M \tag{5.3}$$

onde o_t é o estado observado (novo fonema apresentado ao sistema) e s_k é o símbolo correspondente ao fonema (letra). Por exemplo, caso a palavra *maraca* tenha sido observada e a palavra *macaca* esteja modelada, o fonema /r/ será convertido para /c/. Isto porque, dada a probabilidade da seqüência modelada, é mais provável que o símbolo correto seja *c* e não *r*.

O processamento de um modelo de Markov oculto para reconhecimento de fala dá-se normalmente através de um dos três 3 métodos (TEBELSKIS, 1995):

- uso do algoritmo de avanço (*foward*) para identificação de uma única palavra;
- uso do algoritmo de Viterbi para identificação de um conjunto de palavras (fala contínua);
- uso do algoritmo de avanço-retrocesso (*foward-backward* ou Baum-Welch) para treinamento do modelo.

O algoritmo de avanço define qual é a melhor seqüência de estados existente para dada seqüência observada. Para tanto, são calculadas as probabilidades α para cada estado j num tempo t, com base na matriz A de probabilidades de transição e no vetor B de probabilidades de símbolos observados:

$$\alpha_j(t) = \sum_i \alpha_i(t-1) a_{ij} b_j(s_t) \tag{5.4}$$

Na inicialização do algoritmo, é feito $\alpha_j(0) = 1.0$ para o estado inicial e 0.0 para os demais estados. Para cada instante t de tempo é gerado o estado com maior probabilidade,

de forma a gerar uma seqüência de T estados, os quais podem corresponder a uma palavra, por exemplo.

O algoritmo de Viterbi é uma alternativa viável para o reconhecimento de palavras em fala contínua, por serem infinitas as possibilidades de modelagem através do algoritmo de avanço. O algoritmo de Viterbi realiza uma aproximação da melhor seqüência de símbolos. Assim, ao invés de serem calculadas as probabilidades através do somatório, este é substituído pela função de máximo:

$$\nu_j(t) = \max_i [\nu_i(t-1)a_{ij}b_j(s_t)] \tag{5.5}$$

A funcionalidade deste algoritmo é idêntica ao de avanço, embora deva-se observar que as seqüências T devem variar devido ao diferente tamanho das palavras. Desta forma, o algoritmo deve ser chamado recursivamente para analisar as possíveis derivações descritas na gramática de estados. Aquela seqüência derivada que obtiver maior probabilidade será admitida como a reconhecida.

O algoritmo de avanço-retrocesso é um treinamento dos modelos de Markov segundo as seqüências observadas. O algoritmo de avanço é realizado a partir de seqüências estimadas. Para uma modelagem mais robusta, é necessário realizarem-se os cálculos das probabilidades das seqüências efetivamente ocorridas após as observações.

Para tanto, calcula-se o retrocesso β para a obtenção das probabilidades dos estados já ocorridos, na forma:

$$\beta_j(t) = \sum_i \beta_i(t+1)a_{ji}b_i(s_{t+1}) \tag{5.6}$$

Com base nas probabilidades α anteriormente obtidas no processo de avanço e nas β agora calculadas, é possível a estimação da transição de estados Υ dada seqüência de símbolos s:

$$\Upsilon_{ij}(t) = \frac{\alpha_i(t)a_{ij}b_j(s_{t+1})\beta_i(t+1)}{\sum_k \alpha_k(t)} \tag{5.7}$$

A partir das novas transições de seqüências, as probabilidades podem ser otimizadas através da estimação de novos valores para o modelo. Para tanto, deve ser feita a reestimação das probabilidades de transição do modelo e de símbolos observados.

Para reestimar as transições entre estados calcula-se o número de ocorrências de uma certa transição sobre o número total de ocorrências:

$$\overline{a}_{ij} = P(i \to j) = \frac{N(i \to j)}{N(i \to *)} = \frac{\sum_t \Upsilon_{ij}(t)}{\sum_j \sum_t \Upsilon_{ij}(t)} \tag{5.8}$$

E para nova estimação dos símbolos observados, calcula-se o número de vezes que ocorreu dado símbolo u sobre o total de observações:

$$\overline{b}_{ij} = P(i,u) = \frac{N(i,u)}{N(i)} = \frac{\sum_{t:(y_t=u)} \sum_j \Upsilon_{ij}(t)}{\sum_j \sum_t \Upsilon_{ij}(t)} \tag{5.9}$$

A partir dos novos parâmetros, é possível reaplicar o algoritmo, num refinamento progressivo dos parâmetros, ajustando-os adequadamente para que melhor reconheça futuros novos padrões (fonemas) não-treinados.

Como comentado anteriormente, o objetivo da modelagem de Markov é obter a melhor seqüência possível de fonemas, permitindo a identificação de uma palavra por aproximação estatística. Desta forma, podem ser construídos sistemas híbridos de reconhecimento de fala, onde as características do sinal são adquiridas com redes neurais e a modelagem temporal dos fonemas feita por modelos de Markov.

5.4 Redes Neurais na Análise da Fala

A utilização mais comum de redes neurais tem sido voltadas para o reconhecimento de fonemas para posterior processamento por modelos de Markov, que realizam a análise temporal da composição destes fonemas (MORGAN; BOULARD, 1995). Neste sentido, os modelos Perceptron Multicamadas, como o backpropagation, são largamente utilizados.

Como comentado na seção 5.3, os HMMs são aplicados para a identificação temporal de fonemas, palavras ou frases. Para tanto, devem ser construídos modelos de transições de estados e treinadas as probabilidades destas transições.

Os HMMs podem ser substituídos por redes neurais que aprendem seqüências temporais. Estas não necessitam de modelagem prévia, basta conceber os limites da representação, em número de bits, por exemplo. Os modelos numa rede neural são construídos internamente pela adaptação de seus pesos aos padrões de treinamento (seqüências de fonemas ou palavras).

Na presente Tese, portanto, sustenta-se que redes neurais artificiais podem substituir, satisfatoriamente e de forma subsimbólica (conexionista), os atuais modelos ocultos de Markov e as árvores de parser. Assim, redes recorrentes, tal como a SRN (*Simple Recurrent Network*), redes auto-associativas, como SOM (*Self-Organizing Map*) e suas derivadas TKM (*Temporal Kohonen Map*) e SARDNET (*Sequential Activation Retention and Decay Network*), além de redes recursivas, como a RAAM (*Recursive Auto-Associative Memory*), podem constituir sistemas robustos de parser da linguagem. Segundo Steedman (STEEDMAN, 1999), a combinação de modelos conexionistas seria uma alternativa viável para substituir as atuais modelagens probabilística e simbólica.

5.4.1 Perceptron multicamadas e derivadas

As redes neurais artificiais são amplamente utilizadas para o treinamento e posterior reconhecimento de padrões. Para os sistemas de reconhecimento de fala, os modelos mais utilizados são o backpropagation e as redes recorrentes.

5.4.1.1 Backpropagation

É um modelo cuja arquitetura é composta por três ou mais camadas de perceptrons interconectados, como pode ser visualizado na figura 5.5. Estes perceptrons têm uma diferença fundamental: eles utilizam uma função do tipo *sigmoid* como função de limiar. Esta é uma função não-linear para ampliar o potencial de classificação de um modelo. Essa variação foi o que possibilitou a este e outros modelos realizarem representações complexas. A função sigmoid tem a forma:

$$S_{q_i} = sgm(S_i) = 1/(1 + e^{-(S_i - \alpha)}) \tag{5.10}$$

onde S_{q_i} é a saída quantizada do nodo i, S é a saída resultante da soma ponderada do nodo i e α é o coeficiente de limiar.

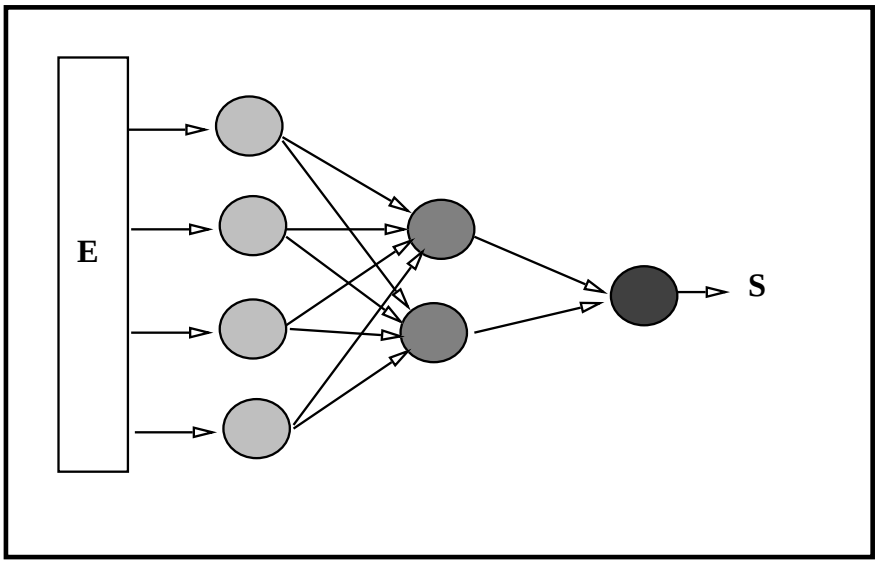

Figura 5.5: Rede neural backpropagation.

Neste modelo, o erro obtido na saída é transferido para as camadas intermediárias. Daí o nome *retropropagação* (backpropagation). Isso se dá pela necessidade de ajuste dos neurônios que não têm contato com a saída, necessitando, assim, de algum parâmetro para atualização dos pesos.

O cálculo do erro começa na última camada, ele tem a forma:

$$\varepsilon s_i(t) = S_{q_i}(t)(1 - S_{q_i}(t))(d_i(t) - S_{q_i}(t)) \tag{5.11}$$

onde S é a saída quantizada, d a saída desejada, e i o nodo atual. A partir deste erro são ajustados os pesos da última camada:

$$P_i(t+1) = P_i(t) + \alpha\varepsilon s_i(t)En(t) \tag{5.12}$$

onde P é o vetor de pesos, α é o coeficiente de aprendizado e En o vetor resultante da saída da camada anterior.

O erro da(s) camada(s) intermediária(s) é feito a partir do erro da camada de saída:

$$\varepsilon_i(t) = En(t)(1 - En(t)) \sum_k \varepsilon_k p_{ik}(t) \tag{5.13}$$

onde En é o vetor resultante da saída da camada anterior até esta camada intermediária; k é o número de nodos conectados a seguir do atual; ε é o erro do nodo k; p é o peso correspondente à conexão do nodo atual com o nodo k. A partir deste erro, são calculados os pesos:

$$P_i(t+1) = Pi(t) + \alpha\varepsilon_i(t)En(t) + \mu(P_i(t) - P_i(t-1)) \tag{5.14}$$

onde μ é um coeficiente de aceleração de convergência denominado *momentum*.

O algoritmo utilizado consiste nos passos:

1. Inicializar os pesos e coeficientes de limiar com valores pequenos e randômicos.
2. Apresentar o vetor de entrada (padrão) e a saída desejada.

3. Calcule a saída linear:
$$S(t) = \sum_{i=0}^{n-1} P_i(t)E_i(t) \tag{5.15}$$

4. Aplique a função sigmoid da equação 5.10.

5. Atualize os pesos da última camada, conforme equação 5.12.

6. e o cálculo do erro da última camada, segundo a equação 5.11.

7. Atualize os pesos da(s) camada(s) intermediária(s), com a equação 5.14 e seu erro pela equação 5.13.

8. Voltar ao passo 2 até que atinja um valor próximo ao da saída desejada.

O algoritmo apresentado é considerado clássico, existindo atualmente versões mais otimizadas de treinamento, tal como counterpropagation e quickpropagation.

5.4.1.2 Redes de atraso de tempo, recorrentes e recursivas

As redes de atraso de tempo e recorrentes foram criadas para o aprendizado de seqüências temporais. Segundo (HERTZ; KROGH; PALMER, 1991) e (HAYKIN, 2001), este tipo de redes tem como objetivo o reconhecimento de seqüências, a reprodução de seqüências ou a associação temporal. O reconhecimento de seqüências tem por objetivo sinalizar a identificação de dada cadeia de elementos. A reprodução permite a complementação de uma seqüência, a partir de seus elementos iniciais. A associação temporal relaciona uma seqüência de entrada com outra de saída. Na presente Tese utilizaremos as redes recorrentes como reconhecedoras de seqüências.

Além das redes recorrentes, a seguir também é apresentada uma rede recursiva, voltada à representação de árvores de parsing, por sua capacidade de compactação de gramáticas. A rede recursiva complementa a recorrente na análise da linguagem, uma vez que é feita a representação subsimbólica da estrutura gramatical na recursiva que posteriormente terá sua seqüência registrada na recorrente. Mais detalhes sobre este processo serão apresentados no capítulo 6.

TDNN

A rede neural por atraso de tempo (TDNN - *Time Delay Neural Network*) foi projetada para simular a simetria do tempo de um espectograma (HAYKIN, 2001). Isso é feito construindo-se uma rede backpropagation, como descrita anteriormente, mas os neurônios da camada de entrada, ou ainda da camada oculta, são replicados *n* vezes para armazenar padrões de entrada anteriores ($z^{-1}, z^{-2}, ...$). Desta forma, a cada apresentação de um padrão (fonema), são também apresentadas simultaneamente os *n* padrões anteriores.

Em sistemas desenvolvidos, comprovou-se que redes TDNN necessitam de menor quantidade de pesos que redes perceptron multicamadas, reduzindo, portanto, a quantidade de memória e o tempo de treinamento. Além disso, as TDNNs têm obtido taxas de reconhecimento superiores ao de redes backpropagation (CASAGRANDE, 1999).

SRN

Uma SRN (*Simple Recurrent Network*) é composta por unidades de contexto armazenam os pesos de uma camada oculta por um passo de tempo, realimentando-os na entrada

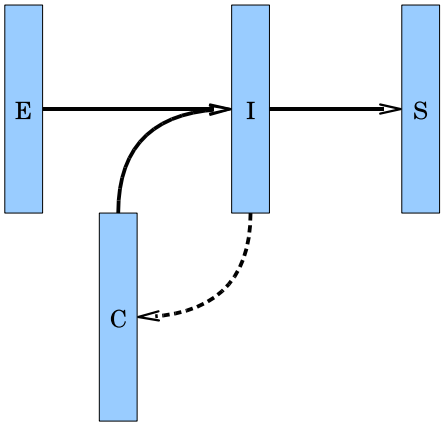

Figura 5.6: Rede neural recorrente simples.

da rede ou de volta à camada oculta. Esta característica permite à rede realizar o armazenamento do estado do sistema, que será utilizado no reforço do aprendizado do próximo estado (HAYKIN, 2001). Wermter e Weber afirmam que este método é mais consistente que modelos de *n-gramas* (ver seção 5.3), que armazenam apenas *n* palavras anteriores, enquanto a rede recorrente possui um limite maior para o armazenamento (WERMTER; WEBER, 1997).

Neste sentido, Elman (ELMAN, 1990) afirma que são necessárias 4 fases no treinamento de uma SRN. Inicia-se por frases simples, passando para simples com algumas complexas, depois com poucas simples e muitas complexas e finalmente com todas complexas. Ao final do processo, a rede é capaz de identificar até mesmo seqüências não treinadas, generalizando o processo de identificação.

Uma SRN nada mais é que um perceptron multicamadas com uma *camada de contexto*. Esta fica juntamente com a camada intermediária, tem o mesmo tamanho, recebe sua saída e devolve o resultado para sua entrada. A camada intermediária, portanto, é alimentada pela entrada e pela camada de contexto, conforme mostra a figura 5.6.

A ativação de um neurônio *i* de saída numa iteração (época) *t* de uma SRN é dada por

$$S_i(t) = \sum_j p_{ij} e_j(t) + b_i + \sum_k p_{ik} c_k(t-1) \tag{5.16}$$

onde p é o peso entre neurônios, e é a entrada, b é o coeficiente *bias* e c é a entrada oriunda da camada de contexto. Deve-se observar que a saída da camada de contexto corresponde à ativação do estado anterior $(t-1)$ para esta camada.

Após é feito o treinamento, realizado normalmente com o ajuste de pesos na forma original do modelo backpropagation. O objetivo é obter na saída a seqüência correta do padrão seguinte ao treinado, ou seja, se for treinada a seqüência *abc*, após apresentado *b* deve-se ter *c* na saída da rede.

Pode-se concluir, portanto, que as redes SRNs podem ser aplicadas ao reconhecimento temporal dos padrões da fala. Caso sejam gerados coeficientes de segmentos de palavras (sílabas), estes poderão (deverão) ser treinados nas SRNs para posterior reconhecimento da seqüência temporal das sílabas ou das palavras que elas compõem.

RAAM

As redes RAAMs, por sua vez, podem substituir as árvores de parser, uma vez que permitem o armazenamento de estruturas hierárquicas binárias (árvores binárias), se-

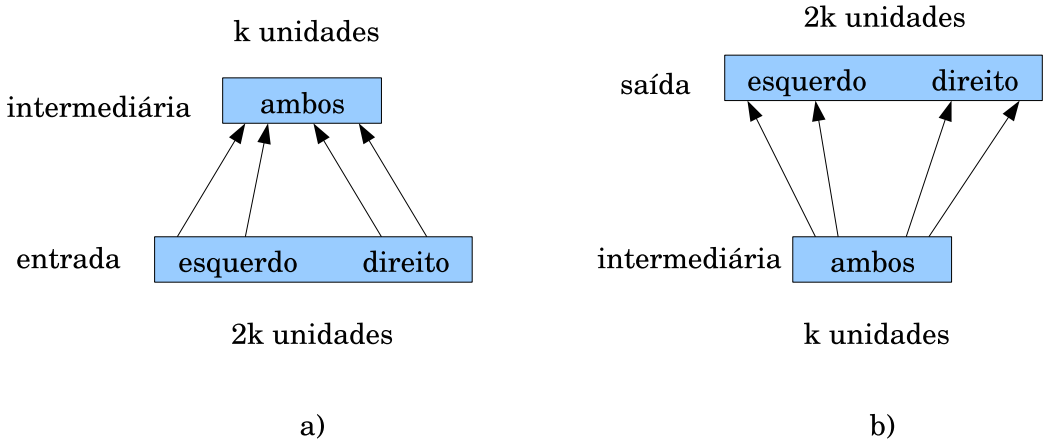

Figura 5.7: Forma de a) codificação e b) decodificação na rede RAAM.

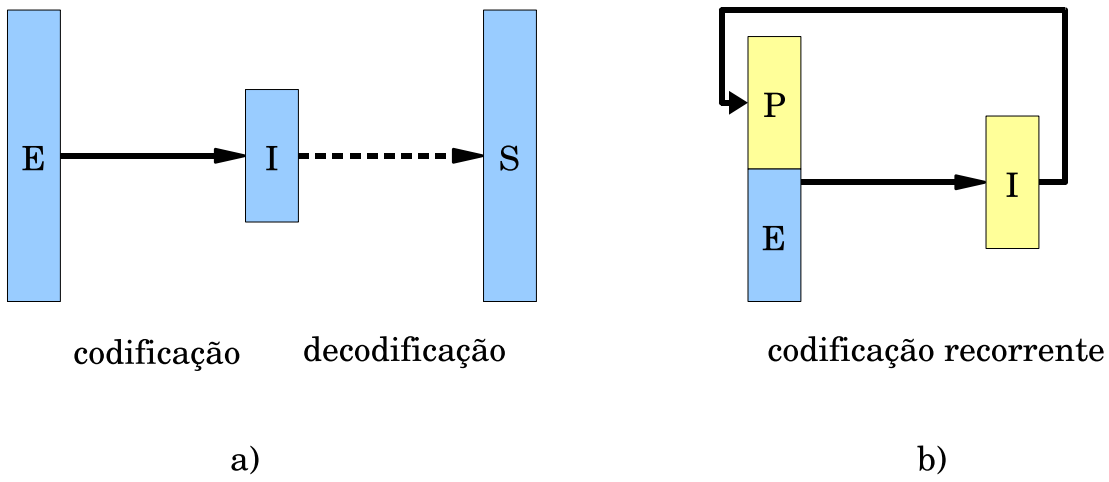

Figura 5.8: Rede neural recorrente simples: a) processos de codificação e decodificação; b) detalhe da codificação por SRAAM.

gundo Pollack, seu criador (POLLACK, 1990). Como qualquer rede neural, a RAAM permite que a sua modelagem interna (pesos) sejam adaptados pelos padrões de treinamento. Assim, observa-se que toda árvore de parser, mesmo as probabilísticas, devem ser previamente modeladas.

A RAAM, assim como a SRN, também é uma rede perceptron multicamadas, mas sem o processo completo, por funcionar como um codificador/decodificador. Na fase de codificação, não é utilizado o sinal de saída para obtenção de respostas. Mas as saídas dos neurônios da camada intermediária são usados recursivamente para compactação de dois (ou mais) termos, um à direita e outro à esquerda, conforme mostra a figura 5.7.

O treinamento da RAAM também segue geralmente o tradicional algoritmo backpropagation. Na fase de decodificação, ocorre o processo inverso, de reconhecimento, mas iniciando com a ativação da camada intermediária e obtendo-se os valores originais na camada de saída. Uma visão destes passos pode ser observada na figura 5.8 a).

Para a realização do armazenamento de uma árvore, onde A e B são filho de R1, e C

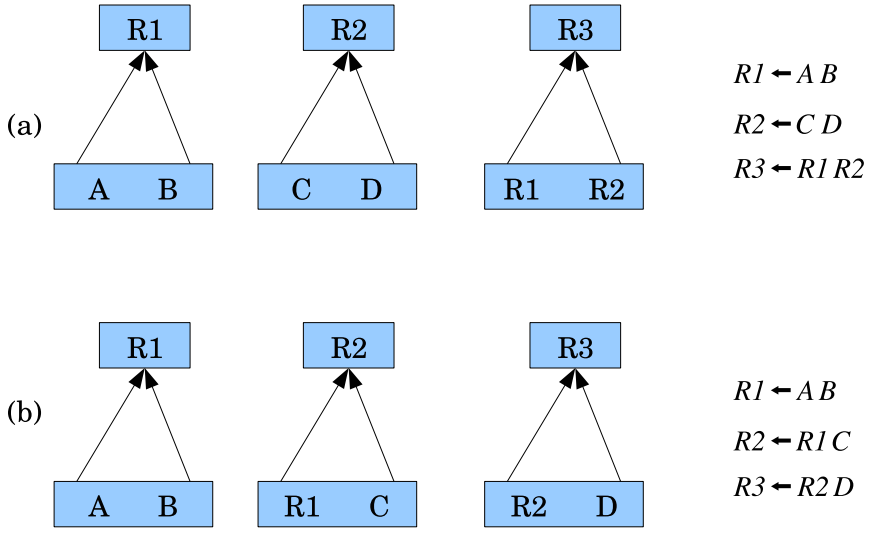

Figura 5.9: Representação de gramáticas na RAAM: a) forma normal ou b) com empilhamento SRAAM.

e D são filhos de R2, treina-se A e B, obtendo-se um resultado R1 na camada escondida, e treina-se outra rede com C e D, com resultado R2. Finalmente, treina-se R1 e R2 para armazenar toda a estrutura, que fica representada em R3 (veja figura 5.9a). Como se pode observar, a rede é a mesma, apenas é necessário o armazenamento em separado das camadas escondidas.

Numa representação funcional, se R1 é (A B) e R2 é (C D), então R3 é equivalente a uma árvore binária ((A B) (C D)). A partir desta premissa observa-se que é possível a construção de estruturas em árvore binária com uma rede neural RAAM.

Além da compactação obtida pelo processamento descrito, também podem ser armazenadas seqüências de padrões. Para tanto, Pollack desenvolveu a SRAAM (Sequential RAAM) (POLLACK, 1990). Esta rede, por exemplo, pode armazenar um padrão vazio mais o A, obtendo um resultado R1. Após, treina-se R1 e B, obtendo R2, e assim sucessivamente (veja figuras 5.8 a) e 5.9b).

Para se ilustrar a capacidade de codificação de uma dada gramática *G*, sejam as seguintes regras (POLLACK, 1990):

```
S -> NP VP | NP V
NP -> D AP | D N | NP PP
PP -> P NP
VP -> V NP | V PP
AP -> A AP | A N
```

Dada a gramática G, é possível o treinamento na RAAM das seguintes seqüências S definidas pelas árvores binárias (em representação funcional):

```
(D (A (A (A N))))
((D N) (P (D N)))
(V (D N))
(P (D (A N)))
((D N) V)
```

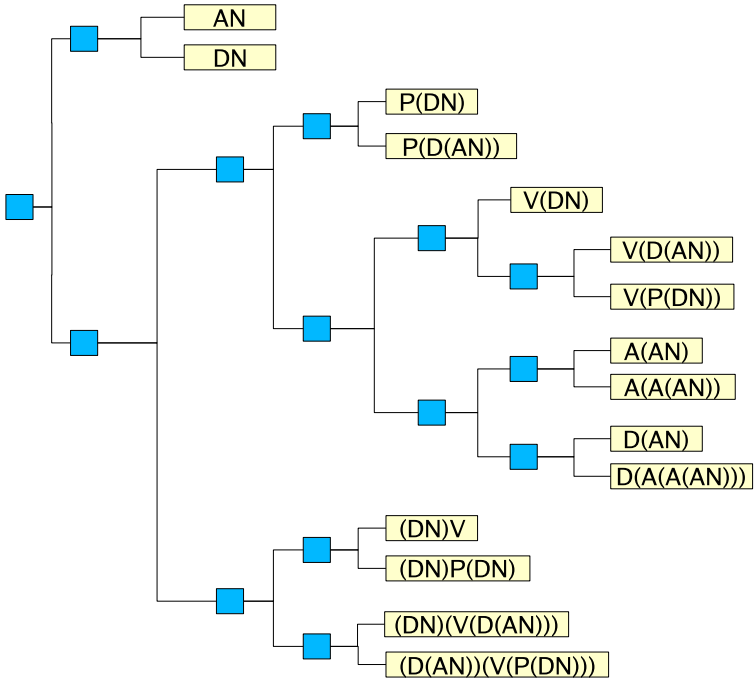

Figura 5.10: Decodificação gerada para a gramática *G*.

```
((D N) (V (D (A N))))
((D (A N)) (V (P (D N))))
```

Aplicando-se o treinamento numa SRN com 20 nodos de entrada, 10 intermediários e 20 de saída, obtém-se a codificação das seqüências desejadas como representado na árvore da figura 5.10 (POLLACK, 1990). Esta árvore é um agrupamento hierárquico utilizado por Pollack para facilitar a visualização dos valores numéricos compactados na camada intermediária da RAAM. Esta representação permite observar a capacidade de estruturação das seqüências S da gramática G proporcionada pela RAAM.

5.4.2 SOM e derivadas

O Mapa Auto-Organizável (SOM - *Self-Organizing Map*), de professor Teuvo Kohonen (KOHONEN, 1984), é utilizado na presente Tese com a intenção de identificar categorias semânticas, como sugerido em (KOHONEN, 1990). A seguir será reproduzida a apresentação do modelo SOM como descrito na Dissertação de Mestrado em (MÜLLER, 1996) e complementado com modelos derivados que são utilizados nesta Tese.

A concepção do modelo SOM foi derivada do estudo da formação de regiões de ativação no cérebro segundo as mesmas funções de percepção, controle e memória (RITTER; MARTINEZ; SCHULTEN, 1992). De forma a simular a auto-associação de estímulos segundo os sinais de ativação, Kohonen propôs um modelo matemático que permite o agrupamento de padrões similares. Este agrupamento se dá em áreas de um mapa formado por neurônios (elementos de processamento), o que permite a identificação de padrões com características comuns.

Representação Topográfica

O modelo do Mapa Auto-Organizável possui duas *camadas* de neurônios, onde a primeira possui tantos neurônios quanto for o tamanho do vetor de valores de entradas, e a

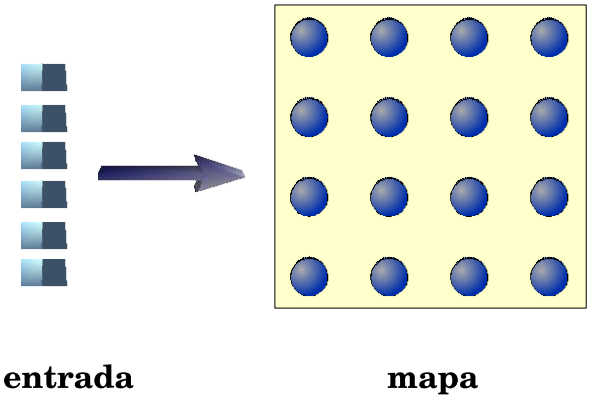

Figura 5.11: Mapa Auto-Organizável.

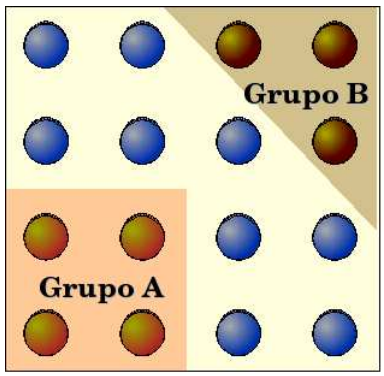

Figura 5.12: Agrupamentos de neurônios proporcionados pelo Mapa Auto-Organizável.

segunda possui o número de neurônios e a forma - retangular ou hexagonal - dependente da aplicação a que se destina a rede. A denominação de *mapa* ocorre devido ao mapeamento de características que é feito nesta segunda camada da rede e segue a representação característica, como mostrada na figura 5.11. Ela forma um plano bidimensional no qual os padrões semelhantes agrupam-se em neurônios próximos, permitindo, assim, grupos por zonas do mapa, conforme vê-se na figura 5.12.

Quanto às conexões, todos os neurônios da primeira camada são conectados aos da segunda, uma vez que o vetor de entradas (neurônios da primeira camada) são comparados aos pesos de cada neurônio da segunda camada. O modelo ainda compreende interações laterais nos neurônios desta segunda camada, que podem ser entendidas como uma influência da saída de um dado neurônio no ajuste dos pesos dos seus vizinhos. Por esta característica, dá-se o nome de *conexão competitiva* a este tipo de organização de rede.

Motivação Matemática

O Mapa Auto-Organizável de Kohonen tem o princípio de representar a freqüência de ativação média de cada neurônio do mapa, mais a interferência dos sinais superpostos dos demais neurônios vizinhos. Este princípio pode ser descrito pelo sistema não-linear de equações:

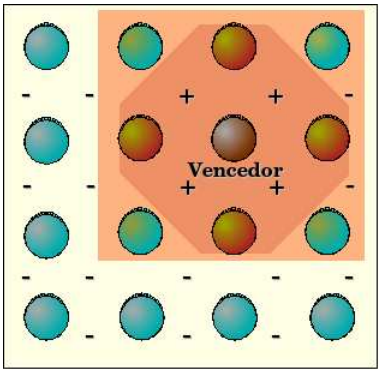

Figura 5.13: Os neurônios próximos recebem sinapses excitatórias e os mais distantes recebem sinais inibitórios.

$$f_r = \mu \left(\sum_l w_{rl}\, v_l + \sum_{r'} g_{rr'} f_{r'} - \theta \right) \tag{5.17}$$

onde f_r é a ativação média para o neurônio r, que envolve a ativação do neurônio, através da soma ponderada do vetor de pesos w_{rl} e vetor de entradas v_l menos o limiar de ativação θ, somados à soma ponderada dos sinais de ativação do neurônio vizinho $f_{r'}$ e a intensidade da sinapse $g_{rr'}$, entre o neurônio atual r e o vizinho r'. O símbolo μ representa uma função do tipo *sigmoid*, que normaliza a resposta de ativação para o intervalo (0;1).

A resposta de ativação do neurônio ($\sum_l w_{rl}\, v_l$) é utilizada em grande parte dos modelos de RNAs, porém Kohonen insere em seu modelo a interferência dos neurônios vizinhos na resposta de ativação de um dado neurônio. Isso se dá através do controle da função $g_{rr'}$ que é definida segundo alguma distribuição probabilística, dado que haja sinapses inibitórias ($g_{rr'} < 0$) para grandes distâncias entre os neurônios e excitatórias ($g_{rr'} > 0$) para pequenas distâncias. Esta distância seria a distância vetorial entre dois pontos, ou seja, $\|r - r'\|$.

O raciocínio anterior permite uma simplificação, afirmando-se que *o neurônio vizinho de maior ativação será aquele com menor distância vetorial entre os pesos e as entradas*, ou seja,

$$\|v - w_{r'}\| = \min_r \|v - w_r\| \tag{5.18}$$

Este cálculo permite uma aproximação do valor de r', necessário para chegarmos às soluções da equação 5.17.

Como para satisfazer-se o cálculo de 5.17 necessita-se ainda a determinação da intensidade dos sinais dos neurônios vizinhos, ou seja, a forma de distribuição de seus sinais, é definida uma função de distribuição $h_{rr'}$ entre dois neurônios. Seguindo o raciocínio da função $g_{rr'}$ de 5.17, a distribuição deve realizar uma sinapse excitatória para os neurônios da vizinhança do neurônio calculado, e inibitória para os neurônios mais distantes, conforme visualiza-se na figura 5.13.

Essa forma de distribuição é possível de ser compreendida como uma troca de pesos sinápticos, na forma:

$$w_r^t = w_r^{t-1} + \epsilon\, h_{rr'}(v - w_r^{t-1}) \tag{5.19}$$

onde a atualização dos pesos do neurônio r, dado por w_r^t, é resultante da aplicação de um valor de adaptação ϵ juntamente com o resultado da função de distribuição $h_{rr'}$ sobre a diferença dos valores de entrada e de pesos $v - w_r^{t-1}$. A função de adaptação para o valor ϵ deve ser decrescente com o tempo t, permitindo a adaptação gradativa à redução da diferença $v - w_r^{t-1}$.

Em outras palavras, esse processo permite a adaptação dos pesos às características do padrão de entrada da rede. Isso se dá através da alteração gradativa dos pesos de blocos de neurônios. Estes blocos são definidos através da função de distribuição $h_{rr'}$, e terão seu tamanho reduzido com o aumento do tempo t. Isso porque cada saída de neurônio tenderá a responder por um dado vetor de entradas. Obviamente não se quer uma rede neural para o aprendizado de apenas um padrão, um vetor de entradas, o que se quer é a representação de diversos padrões, tantos quantos forem possíveis de serem adaptados na rede. Assim, cada padrão de entrada será melhor adaptado por um determinado neurônio. A ordenação dos neurônios que dão a resposta dá-se de forma a representar as proximidades numéricas dos padrões aprendidos ($v - w_r^{t-1}$), e, por conseqüência, a proximidade entre os neurônios com uma resposta semelhante ($\|r - r'\|$).

Como resultado desse processo, os padrões com codificação semelhante serão representados por neurônios próximos, dado um plano bidimensional. Isso permite a criação de zonas de características que são descritas pelos próprios padrões de entrada. Compreende-se, então, um processo de auto-organização de características, ou seja, o modelo que Kohonen propôs é um sistema de auto-classificação de padrões, que registra as características com as quais foram codificados os vetores de entrada. Por estas propriedades, diz-se que este é um modelo com *auto-aprendizado*.

Algoritmo SOM

Para se ter uma idéia mais clara da motivação matemática, a seguir é apresentado o algoritmo básico do modelo de Kohonen, com variantes propostas pelo próprio Kohonen em aplicações de mapas semânticos (KOHONEN, 1990), utilizado na implementação computacional do presente trabalho.

1. Inicializar os pesos da rede com valores randômicos no intervalo [0.01;0.1].

2. Inserir o padrão de entrada.

3. Calcular as distâncias vetoriais ($\|v - w_r\|$):

$$d_l = \sqrt{\sum_{i=0}^{N-1} (v_l(t) - w_{rl}(t))^2} \tag{5.20}$$

onde d_l é a distância entre a saída do nodo j com a entrada e N é o número de entradas.

1. Selecionar a menor distância, segundo a equação 5.18.

2. Atualizar os pesos, segundo a equação 5.19, com uma distribuição gaussiana para $h_{rr'}$:

$$h_{rr'} = exp\left(-\frac{\|r - r'\|^2}{\sigma^2}\right) \tag{5.21}$$

onde o raio σ é decrementado com o tempo:

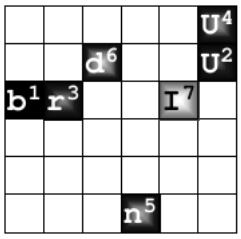

Figura 5.14: Seqüência de reconhecimento dos fonemas da palavra *burundi* na rede SARDNET (JAMES; MIIKKULAINEN, 1995).

$$\sigma(t) = \sigma_i \left(\frac{\sigma_i}{\sigma_f} \right)^{\frac{t}{t_{max}}} \tag{5.22}$$

onde σ_i é o valor inicial do raio e σ_f seu valor final.

1. Repetir a partir do passo 2, até t_{max}.

Os procedimentos apresentados formam o módulo central do protótipo computacional, onde são calculados os resultados dos mapas de palavras e de frases. Os demais módulos do protótipo são destinados ao reconhecimento de padrões e ao pré-processamento das palavras e frases.

5.4.2.1 SARDNET

Um trabalho semelhante ao de Elman foi realizado por James e Miikkulainen ao basearem-se no *Temporal Kohonen Map* (TKM) de Chappel e Taylor (1993) e desenvolverem o *Sequential Activation Retention and Decay Network* (SARDNET). Eles utilizaram um mapa de características para treinar a ordem das letras de palavras. Isso permitiu a classificação de um conjunto de categorias de palavras num único mapa de Kohonen (JAMES; MIIKKULAINEN, 1995).

A SARDNET diferencia-se do modelo SOM pela capacidade de armazenar uma seqüência de neurônios vencedores. Dada esta característica, ele permite a identificação das ativações no tempo.

Segundo (JAMES; MIIKKULAINEN, 1995), o algoritmo do modelo SOM seria alterado para a seguinte seqüência:

1. Encontre o vetor de pesos com menor distância com a entrada;
2. assuma 1.0 como ativação para aquela unidade;
3. ajuste os pesos do neurônios vizinhos;
4. exclua o vencedor atual das competições seguintes;
5. decremente os valores de ativação dos demais nodos.

Como resultado deste processo, tem-se o treinamento de uma seqüência de ativação de neurônios. Alguns testes apresentados por (JAMES; MIIKKULAINEN, 1995) mostraram o reconhecimento temporal de palavras de três sílabas com taxas acima de 97% de sucesso. Um dos mapas treinados por James está reproduzido na figura 5.14.

5.5 Ferramentas para análise da linguagem falada

Neste capítulo foram analisadas técnicas de processamento de linguagem natural normalmente usadas em textos. Após, passou-se ao estudo de técnicas para processamento de linguagem falada, no sentido do escopo da presente Tese, uma vez que a análise textual não é suficiente neste contexto.

A partir dos paradigmas de análise textual descritos são realizadas mudanças na forma de processamento para sua adaptação à fala. As ferramentas de análise da linguagem falada são basicamente estocásticas, que realizam a passagem da representação da fala para a escrita, ou seja, a tradução da linguagem numérica da codificação do sinal para os símbolos gráficos.

Como visto, o HMM é uma técnica largamente utilizada para estimação dos símbolos indicados pelo sinal codificado. A sua principal característica é a capacidade da modelagem temporal de seqüências de símbolos. Apesar de suas propriedades, ela não será necessária para a implementação do modelo COMFALA.

As redes neurais analisadas neste capítulo darão suporte às análises sintática e semântica no modelo computacional. Para análise sintática procura-se estabelecer o registro hierárquico e temporal da linguagem, o que pode ser estabelecido com as redes RAAM, SRN e SARDNET, como posteriormente descrito na seção 6.2.2. Para análise semântica é necessária a definição dos contextos de construção das frases, agrupamentos estes que podem ser discernidos com redes SOM. No capítulo seguinte estão descritos como as ondeletas e redes neurais podem ser compostas para a definição da implementação do modelo COMFALA.

6 MODELO COMPUTACIONAL *COMFALA*

A experiência de sistemas em Inteligência Artificial (IA) mostra que muitas vezes a inspiração biológica, além da cognitiva, torna possível uma modelagem computacional mais robusta dos processos naturais que se deseja representar. Como aqui está sendo tratado do processo de audição humana e da compreensão da linguagem contida nesta ação, torna-se uma exigência realizar a implementação computacional como uma comparação com o processo natural.

No capítulo 2 foram analisados diversos sistemas que envolvem o processamento da linguagem falada, porém nenhum deles apresentou um parâmetro com o processamento biológico. Todos eles foram construídos através de composições de técnicas continuamente evoluídas, mas não foi possível a observação de fundamentos no processamento natural da audição para concepção dos sistemas. Obviamente a reprodução do processamento natural não é uma condição *sine qua non* para a IA, embora a motivação biológica seja um elemento qualificador durante a concepção de um modelo computacional. A motivação biológica permite o esclarecimento da origem e organização das técnicas aplicadas no desenvolvimento de um sistema de IA.

Como introduzido no capítulo 1, a base para a concepção computacional aqui apresentada está no modelo neurocognitivo de audição de frases de Friederici, descrito no capítulo 3. A partir das fases constatadas na investigação natural serão propostas as fases correspondentes no modelo computacional. Os módulos componentes do modelo proposto serão descritos na seção 6.1 e uma proposta para sua implementação encontra-se na seção 6.2.

6.1 Módulos

A presente Tese utiliza o Modelo Neurocognitivo de Processamento da Audição de Frases (MNPAF) com a extensão da fase prosódica, como descrito na seção 3.1, para concepção do Modelo Computacional de Compreensão de Fala (COMFALA), ou em inglês, *Speech Understanding Model* (SUM). Com sua primeira referência ao COMFALA publicada em (MÜLLER; NAVAUX, 2004), o COMFALA é inspirado nas fases do MNPAF, acrescido da nova fase prosódica indicada na seção 3.1.5. Como representado na figura 6.1, o COMFALA é composto dos módulos de *Processamento do sinal*, *Processamento sintático*, *Processamento prosódico-semântico* e *Avaliação*.

Há, nesta Tese, o interesse de reproduzir funcionalmente o MNPAF, sem a preocupação de realizar uma imitação dos processos biológicos ou de sua relação temporal. O MNPAF serve, em suma, como o delineador da estrutura do COMFALA.

Desta forma, o módulo *processamento do sinal* representa a fase 0 do MNPAF por ser o responsável pela segmentação fonológica e o seqüenciamento de fonemas e expressões

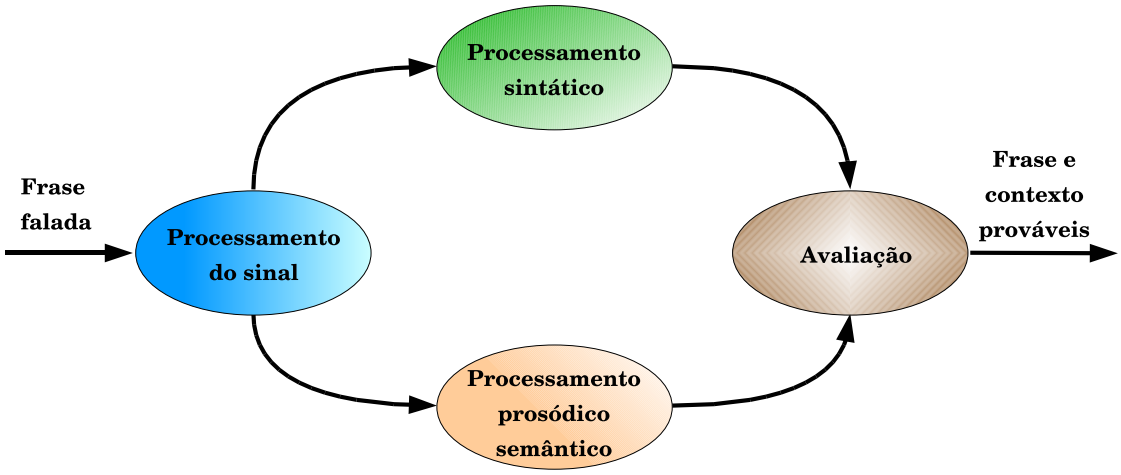

Figura 6.1: Modelo Computacional de Compreensão de Fala.

de fala. O *processamento sintático* segue a idéia da fase 1 por realizar a construção da estrutura sintática. O *processamento prosódico-semântico* corresponde à fase 2 como sendo o relacionamento das características semânticas, acrescido da fase prosódica. Por fim, o módulo *avaliação* equivale à fase 3 por realizar a integração das informações necessárias para estimação do que representa o sinal processado.

6.1.1 Processamento do sinal

Na seqüência de processamento do COMFALA, o sinal de áudio é recebido pelo módulo *processamento do sinal*, onde é segmentado. As partes resultantes são codificadas para posterior reconhecimento dos eventos de comunicação contidos nessas partes. Para o reconhecimento destes eventos há dois caminhos, de processamento distintos e independentes, que se encontrarão novamente na fase de avaliação. Um caminho leva à identificação de padrões lexicais (palavras) e sua organização. O outro caminho leva ao reconhecimento dos padrões não-lexicais (prosódia) e ao contexto de uso dos padrões lexicais (semântica).

Como analisado na seção 3.1.2, o pitch, constituído a partir da onda fundamental (F0), é um elemento fundamental para a segmentação das palavras e para a definição de contextos. O módulo *processamento do sinal* do COMFALA deve, portanto, realizar um tratamento do pitch, de forma a extrair informações tanto lexicais como prosódicas, que auxiliarão na definição das palavras e do contexto semântico associado à entonação.

A implementação do módulo *processamento do sinal* pode ser feita pelas técnicas descritas no capítulo 4. Muitos trabalhos atualmente descritos para reconhecimento de voz realizam a codificação pela extração dos coeficientes cepstrais, na forma analisada na seção 4.2. Por outro lado, outra metodologia mais recente que vem se disseminando rapidamente é a extração de coeficientes por transformadas ondeletas, descrita na seção 4.3.

6.1.2 Processamento sintático

O módulo *processamento sintático* é isolado do *prosódico-semântico* com base no MNPAF. Friederici, ao descrever o MNPAF, constatou que não há interferência entre as áreas cerebrais responsáveis pela sintaxe e pela semântica, a não ser na fase de integra-

ção, ao final do processo. Com base nestas constatações neurocognitivas, o COMFALA descreve o processamento dos módulos *sintático* e *prosódico-semântico* como sendo paralelos e independentes.

As características lexicais do sinal da fala, identificadas no processamento do sinal, são passadas ao módulo *processamento sintático* para organização temporal e lingüística. Sobre o processamento sintático, o MNPAF constata que há etapas de análise bem definidas, como visto na seção 3.1.2, que podem ser transportadas para o COMFALA. Estas etapas seriam:

1. definição da categoria sintática da palavra;

2. compatibilização das categorias na estrutura sintática;

3. caso a estrutura seja coerente, passa-se para a análise semântica - caso contrário a audição deve ser repetida.

Assim, o módulo *processamento sintático* deve possuir padrões de estrutura para os quais as palavras identificadas possam ser encaixadas. Caso a seqüência apresentada não seja reconhecida, todo o processo deve ser repetido com a captura de nova fala. Aqui cabe ressaltar a importância da análise dos padrões temporais das frases. Caso a organização temporal não esteja condizente aos padrões conhecidos, como por exemplo *no coloquei biscoito os pote*, a frase deve ser rejeitada.

A análise de concordância, por outro lado, não provoca rejeição da frase. Com as mesmas palavras do exemplo anterior, mas organizadas temporalmente de forma correta, tem-se *coloquei os biscoito no pote*. Esta frase é estruturalmente aceitável, embora haja um erro de concordância de número entre *os* e *biscoito*.

Por fim, caso haja ambigüidade, como em *coloquei o dinheiro no banco*, o conceito de banco vai depender do contexto de uso que estão inseridas as frases. Se o uso estiver associado a uma casa, banco terá o sentido de um móvel, caso estejam associados a uma agência bancária, o contexto será outro. Os contextos que estarão associados aos termos serão analisados no módulo processamento prosódico-semântico.

O módulo *processamento sintático* pode, a rigor, ser implementado por qualquer sistema de processamento de língua natural (seção 5.1), embora necessite de técnicas de conversão da codificação oriunda do sinal de fala para a forma de palavras. A forma atualmente usada de conversão tem sido através de modelos estocásticos e conexionistas, como os abordados na seção 5.2. Estes sistemas permitem a utilização direta dos coeficientes extraídos pelo processamento do sinal de fala e realizar seu uso como símbolos no processamento sintático.

6.1.3 Processamento prosódico-semântico

O módulo processamento prosódico-semântico visa a resolução de ambigüidades, concordância e gênero, usando contextos de uso que associem corretamente os conceitos envolvidos na frase em análise. Como observado na seção 3.1.3, o MNPAF indica como condição para o processamento semântico a correta construção da frase. O COMFALA, em contrapartida, não foi concebido para que a semântica seja condicionada a uma estruturação sintática correta. Ao contrário, foi observado como processos completamente separados.

Outra característica que o COMFALA não segue rigorosamente o MNPAF é quanto à prosódia. Como o MNPAF não aborda em detalhes a ação da prosódia, foram buscados

trabalhos de outros pesquisadores com relação à ação da prosódia sobre a audição de frases.

Neste sentido, como analisado na seção 3.1.5, não foram constatadas relações cerebrais entre a prosódia e a análise sintática. Por outro lado, com a análise semântica, ao contrário, foi observado um processamento integrado com a prosódia. Segundo as pesquisas de Annett Schirmer e sua equipe, a prosódia emocional auxilia na contextualização de palavras similares e na seleção de contextos (SCHIRMER; KOTZ; FRIEDERICI, 2002; SCHIRMER et al., 2004; SCHIRMER; KOTZ; FRIEDERICI, 2005).

Partindo da premissa de que há uma relação entre a prosódia e a semântica no processamento cerebral, o COMFALA procura reproduzir esta cooperação observada. Desta forma, a prosódia deve atuar no sentido de auxiliar na determinação de contextos semânticos. Não se pode afirmar que a prosódia é o único fator a influenciar na definição de contextos, mas que possui um papel importante neste processo.

Para a implementação deste módulo processamento prosódico-semântico há sistemas que usam a prosódia na semântica, como comentado na seção 2.4.2. Estes sistemas utilizam a prosódia para a definição de contextos de análise da linguagem.

6.1.4 Avaliação

Os resultados da análise sintática e prosódico-semântica que o módulo *avaliação* recebe podem ser analisados por vários ângulos, dependendo do objetivo que se deseja implementar. Ele pode ter uma conotação de sistema de correção, caso o objetivo seja obter uma verificação da construção da frase falada. Pode ter ainda a função de identificação de uma frase ou de um contexto de fala ou de ambos simultaneamente.

Na seção 3.1.4 foram expostas as hipóteses de processamento cerebral relativos à integração das fases anteriores. Como visto, aparentemente ocorrem dois tipos de fenômeno: tentativa de correção de estrutura e de argumento. A correção de estrutura indica a necessidade de reorganização da seqüência temporal de apresentação das palavras. A correção de argumento indica a necessidade de substituição de uma palavra por outra mais coerente com a estrutura. No COMFALA, o módulo avaliação pode ter sua implementação orientada, como citado, para a correção de estrutura e argumento da frase, seguindo o modelo neurocognitivo.

Por outro lado, o MNPAF também aponta a interação de outras áreas cerebrais nesta fase de integração. Fenômenos como a leitura (visão) e escrita (movimento) também podem atuar neste momento, além de outras áreas sensoriais. Uma implementação do COMFALA poderia, portanto, utilizar dados provenientes de sensores de ambiente e outros sistemas para compor a avaliação necessária para a compreensão mais correta possível da fala, complementando as análises sintática e prosódico-semântica.

6.2 Implementação do modelo

O COMFALA é um modelo computacional ao qual podem ser agregadas diversas tecnologias para sua implementação. Com o objetivo de demonstrar sua viabilidade, aqui são descritas algumas técnicas que foram compostas na forma de um protótipo, constituindo um sistema de compreensão de fala.

O protótipo de implementação do modelo está descrito nas seções a seguir. Para o módulo *processamento do sinal* é proposto o uso de ondeletas em sistema Matlab construído pelo autor. Este módulo recebe um sinal da frase falada FF e aplica um algoritmo f_F para obtenção dos coeficientes fonéticos CF:

$$FF \overset{f_F}{\Rightarrow} CF \tag{6.1}$$

e um algoritmo f_P para obtenção dos coeficientes prosódicos CP:

$$FF \overset{f_P}{\Rightarrow} CP \tag{6.2}$$

Para o *processamento sintático*, tem-se o sistema conexionista SARDSRN-RAAM, descrito em (MAYBERRY III; MIIKKULAINEN, 1999). Este sistema usará os coeficientes fonéticos CF como padrões de entrada, para uma análise sintática ANS, obtendo-se uma frase sintaticamente similar FSS:

$$CF \overset{ANS}{\Rightarrow} FSS \tag{6.3}$$

Para o *processamento prosódico-semântico* é usado um sistema de mapas SOM desenvolvido pelo autor. Como entrada deste processamento tem-se a composição dos CF e CP numa representação da fala RF:

$$RF = CF \cup CP \tag{6.4}$$

para obtenção do contexto de frase mais provável CFP através da análise semântico-prosódica ASP:

$$RF \overset{ASP}{\Rightarrow} CFP \tag{6.5}$$

Como *avaliação*, há ponderações das saídas sintática e prosódico-semântica visando a aproximação textual e conceitual da frase falada. Assim, são comparadas as respostas FSS e CFP ponderando-se qual obteve melhor aproximação por probabilidade, obtendo-se a frase e/ou contexto prováveis do sinal de fala FCP:

$$FCP = ponderado(FSS; CFP) \tag{6.6}$$

onde $ponderado$ é um algoritmo de comparação das probabilidades obtidas de FSS e CFP com relação aos padrões conhecidos nas análise da linguagem.

A figura 6.1 pode então ser refeita com as definições dos módulos componentes, resultando na figura 6.2.

6.2.1 Processamento do sinal de fala por ondeletas

Com base na análise realizada na seção 4.3.1, a implementação do módulo *processamento do sinal* será realizada utilizando a transformada ondeletas para a extração das características do sinal. Por representar uma análise de domínio tempo-freqüência do sinal, a transformada ondeletas impede a perda ou sobreposição de dados, como ocorre na transformada de Fourier que realiza apenas a análise de domínio freqüência.

Outra razão para o uso de ondeletas é devido à sua resposta de freqüência, que permite um comportamento correspondente ao da cóclea. Como analisado na seção 4.3.3.1, a cóclea funciona como um banco de filtros, e possui uma resposta de freqüência linear até 500 Hz e logarítmica acima de 800 Hz (YANG; WANG; SHAMMA, 1992; GOLDSTEIN, 1994).

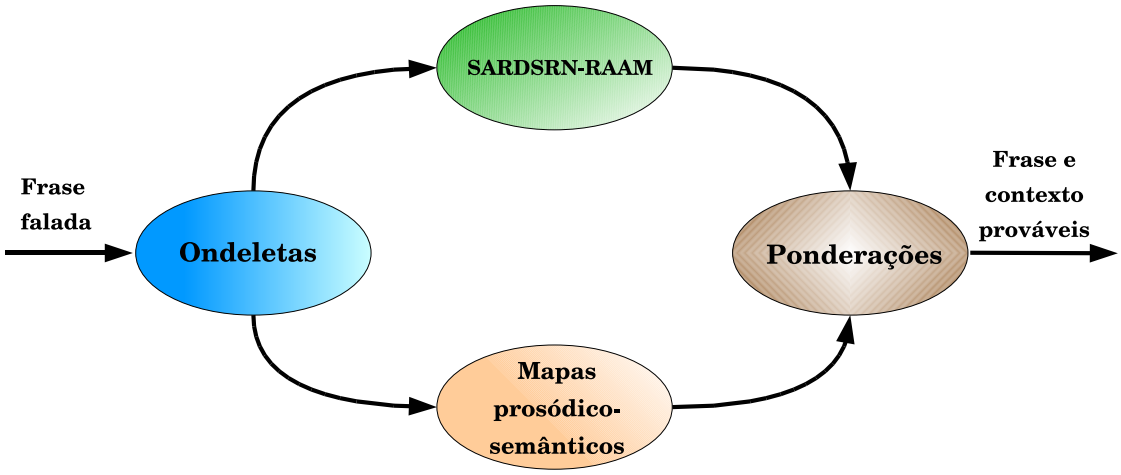

Figura 6.2: Módulos de implementação do COMFALA.

Através das ondeletas serão extraídos os sinais lexicais, que contém palavras e são aqui chamados de *fonéticos*, e os sinais não-lexicais, frutos da análise da onda fundamental (F0), chamados *prosódicos*. Os coeficientes fonéticos obtidos serão usados no processamento sintático e prosódico-semântico. Os coeficientes prosódicos serão usados para análise prosódica-semântica.

6.2.1.1 Coeficientes fonéticos

Segundo o *Modelo de Processamento da Audição* de Yang, descrito na seção 4.3.3.1, é possível a reprodução das fases de análise, que é uma funcionalidade da cóclea, realizando a divisão do sinal em sub-bandas, e a redução, que permite a extração das características de cada banda. Segundo (YANG; WANG; SHAMMA, 1992), as ondeletas podem ser usadas como banco de filtros, obtendo as sub-bandas e para cada uma destas é feita a análise espectral, para a qual pode ser usada a transformada de Fourier.

Conforme discutido na seção 4.3.3.2, diversas pesquisas indicam o uso de um processamento semelhante ao modelo de Yang, ao invés do algoritmo MFCC. Os problemas encontrados no uso do MFCC são, segundo (TUFEKCI; GOWDY, 2000):

- o uso da derivada do cosseno proporciona uma distorção de representação do sinal, uma vez que o ruído em uma faixa de banda afeta a codificação do sinal nas demais faixas;
- um quadro de análise pode conter mais de um fonema, uma vez que os fonemas de alta freqüência podem se sobrepor aos de baixa freqüência.

A análise de multiresolução de ondeletas, descrita na seção 4.3.2, tem sido largamente utilizada como banco de filtros para extração de características do sinal, por sua divisão em sub-bandas (CARNERO; DRYGAJLO, 1999; GUPTA; GILBERT, 2001; AVCI; TURKOGLU; POYRAZ, 2005). Em cada sub-banda extrai-se então a densidade espectral como identificadores de um fonema (KIM; YOUN; LEE, 2000; FAROOQ; DATTA, 2004; RICOTTI, 2005).

A implementação deste processo, descrito na figura 6.3, foi realizada no ambiente Matlab, utilizando as funções relativas à análise de multiresolução em ondeletas. Um arquivo contendo o sinal de fala é lido e codificado em sub-bandas.

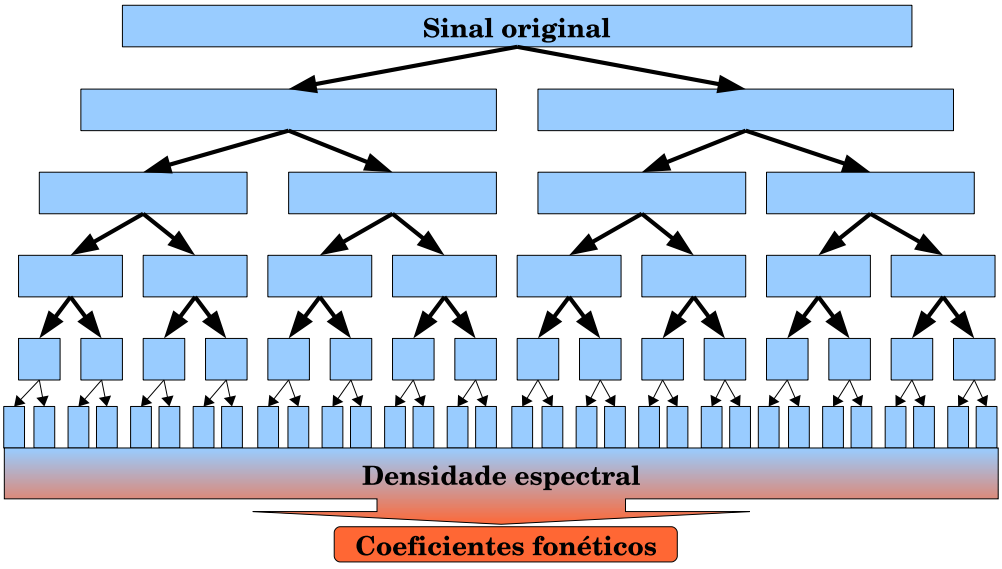

Figura 6.3: Obtenção dos coeficientes fonéticos por análise de multiresolução.

Cada sub-banda tem sua densidade espectral calculada e seu conjunto é usado como coeficientes fonéticos de identificação do sinal de fala apresentado. A densidade espectral D em determinada sub-banda é calculada por

$$D = \frac{\Sigma c^2}{n} \tag{6.7}$$

onde c é o coeficiente da sub-banda e n o número de coeficientes obtido. Para efeito de diminuir a variância entre os coeficientes de uma mesma sub-banda e assim permitir um melhor desempenho nas redes neurais, utiliza-se uma normalização D_N através do logaritmo natural da densidade espectral:

$$D_N = \ln(D) \tag{6.8}$$

O conjunto das normalizações D_N serão o padrão fonético de identificação do que é falado, ou seja, os coeficientes fonéticos CF são definidos como a união de todas as densidades normalizadas obtidas:

$$CF = D_{N_1} \cup \ldots D_{N_i} \cup \ldots D_{N_{nt}} \tag{6.9}$$

onde i é o i-ésimo coeficientes obtido e nt é o número total de folhas da árvore de multiresolução. Esta composição torna-se um padrão que é posteriormente associado a um fonema, palavra ou expressão de forma a representar a fala.

6.2.1.2 Coeficientes prosódicos

Como analisado na seção 4.4.4, o pitch, usado na análise prosódica, pode ser estimado através de ondeletas. Segundo (KADAMBE; BOUDREAUX-BARTELS, 1990, 1992), as ondeletas descrevem os pontos de máximo local, representando uma variação no fluxo da fala. A distância entre dois pontos de máximo representam um instante de fechamento de glote (GCI), o qual permite a identificação de um ciclo da onda fundamental, permitindo a estimação do pitch.

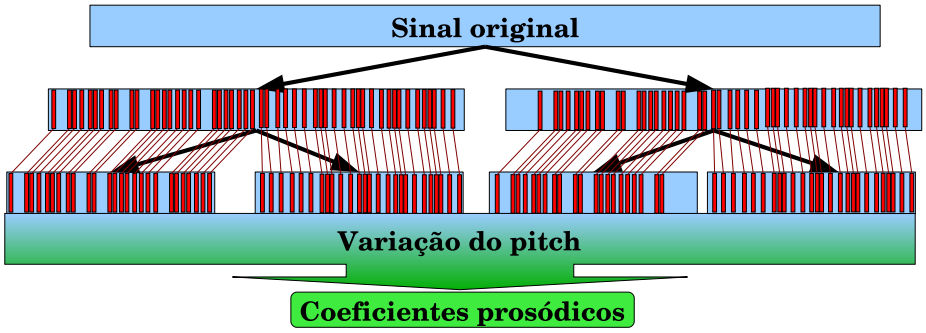

Figura 6.4: Obtenção dos coeficientes prosódicos por análise de multiresolução.

Para confirmação dos pontos de máximo, Kadambe sugere a obtenção destes pontos em duas escalas de análise de ondeletas. Usando-se a análise multiresolução, é possível a obtenção e comparação da ocorrência destes pontos nas duas escalas. Segundo Kadambe, não é necessário descer mais que 3 escalas para obtenção dos dados, mas deve-se observar que o ponto de máximo da escala seguinte deve ser superior a 80% do máximo da escala atual (KADAMBE; BOUDREAUX-BARTELS, 1992). Uma vez obtidos os pontos, é necessário o cálculo de sua freqüência, que será a estimação do pitch desejada. As variações do pitch serão a informação prosódica necessária ao processamento prosódico do COMFALA.

Os coeficientes prosódicos CP são, portanto, os pontos de máximo PM entre os coeficientes de ondeletas obtidos pela transformada T_O de uma escala E e de sua sucessora:

$$CP = PM(T_{O_E}; T_{O_{E+1}}) \tag{6.10}$$

ou seja, CP conterá os maiores valores de coeficientes comparados entre os correspondentes de duas escalas sucessivas.

A implementação da obtenção dos coeficientes prosódicos deu-se também utilizando-se o ambiente Matlab. O arquivo com a fala é lido e os pontos de máximo das segunda e terceira escalas são comparados. Os pontos coincidentes são usados para o cálculo da freqüência da F0. Estes valores de freqüência são a identificação da prosódia do sinal em análise. Este processo está descrito na figura 6.4.

6.2.2 Processamento sintático com o sistema SARDSRN-RAAM

O processamento sintático pode ser resolvido por uma grande quantidade de sistemas existentes, com a condição de que permitam a utilização dos dados numéricos resultantes do processamento do sinal de fala. Para atender à necessidade de implementação deste módulo, foi proposto para esta Tese como critério de escolha um sistema que exigisse uma interferência mínima do usuário para o projeto do parser.

Para atender as exigência de entrada por valores e com projeto semi-automático, poderiam ser usados parsers probabilísticos ou conexionistas. Os probabilísticos usam redes bayesianas, enquanto os conexionistas usam Redes Neurais Artificiais (RNAs), como comentado na seção 5.2. Dentre ambos, as RNAs necessitam menor interferência do usuário, uma vez que são treinadas a partir de exemplos. Com base nesta definição, foram procurados sistemas de parser conexionistas que atendessem à necessidade de implementação do processamento sintático.

Dentre os sistemas pesquisados, foi escolhido o SARDSRN-RAAM, proposto por (MAYBERRY III; MIIKKULAINEN, 1999). Neste sistema, a rede SRN e a SARDNET

juntamente com a RAAM formam um *shift-reduce parser* (MAYBERRY III; MIIKKULAINEN, 1999). Esta forma de parsing refere-se a como é feita a avaliação de uma gramática. O *shift* é a colocação de um termo a avaliar numa pilha e o *reduce* é quando os elementos do topo da pilha recebem uma só representação. Por exemplo, ao processar uma regra $F \rightarrow SN\ SV$ o símbolo F contém uma representação dos elementos empilhados SN e SV.

Conforme analisado na seção 5.4.1.2, a RAAM possui a propriedade de codificar os termos de uma gramática, como o processo *reduce* e também de realizar o empilhamento, ou seja, o *shift*. Desta forma, a RAAM é a responsável pela representação da gramática a ser avaliada pelo sistema SARDSRN-RAAM.

Na seção 5.4.1.2 foi comentado que a SRN tem por característica a previsão do próximo estado de uma seqüência. Infelizmente, segundo (MAYBERRY III; MIIKKULAINEN, 1999), há uma perda de capacidade de memória quando do processamento de longas seqüências. Para solucionar esta limitação Mayberry acrescentou uma SARDNET como mais uma camada intermediária no SRN.

A SARDNET, como visto na seção 5.4.2.1, armazena uma seqüência de neurônios vencedores, permitindo assim o registro temporal de uma cadeia de entradas na rede. Segundo (MAYBERRY III; MIIKKULAINEN, 1999), o acréscimo da SARDNET no SRN incrementou a capacidade de registro temporal, permitindo a identificação de seqüências em frases de tamanho usado na realidade.

6.2.2.1 Características da implementação do SARDSRN-RAAM

O SARDSRN-RAAM é um sistema implementado em linguagem C, mas que utiliza procedimentos gráficos das bibliotecas TCL, TK e BLT, disponíveis para sistemas Unix e derivados. O sistema é de livre uso para fins de ensino e pesquisa, com os direitos reservados a Marshall R. Mayberry. Os códigos fonte do sistema podem ser obtidos através da internet no endereço `http://www.cs.utexas.edu/users/nn/pages/software/abstracts.html#mir`.

Como descrito, na implementação do SARDSRN-RAAM há três redes neurais envolvidas, RAAM, SRN e SARDNET. A atual modelagem da RAAM tem 128 entradas, 64 neurônios na camada intermediária e 128 na saída. Já o SARDSRN possui 64 entradas, sendo 200 neurônios nas camadas intermediárias da SRN e 144 no mapa da SARDNET, e a saída com 64 neurônios.

6.2.2.2 Processamento do SARDSRN-RAAM

O processamento do parsing inicia com a codificação das palavras através do RAAM, segue com o seqüenciamento das palavras na frase composta com o SARDSRN, o qual gera, em sua saída, a representação que é decodificada novamente pela RAAM, a qual indica a frase reconhecida. Para o seqüenciamento temporal das palavras é usado o sistema SARDSRN, onde a camada de entrada é distribuída para uma camada intermediária e para uma rede SARDNET. Esta, por sua vez, também alimenta a camada intermediária. Em paralelo à camada intermediária há uma camada de contexto, o que caracteriza o SRN no SARDSRN. A figura 6.5 mostra este processo, com a entrada do padrão no SARDSRN após a codificação RAAM, passando ao SRN, que é indicado pelos retângulos circulados, e o Sardnet. Depois da saída do SRN, o padrão é decodificado pela rede RAAM.

O primeiro passo para o treinamento do SARDSRN-RAAM é a codificação do léxico pela RAAM. Os coeficientes fonéticos, que representam as palavras obtidas no processamento do sinal, são definidos como os códigos que representam os terminais da gramática

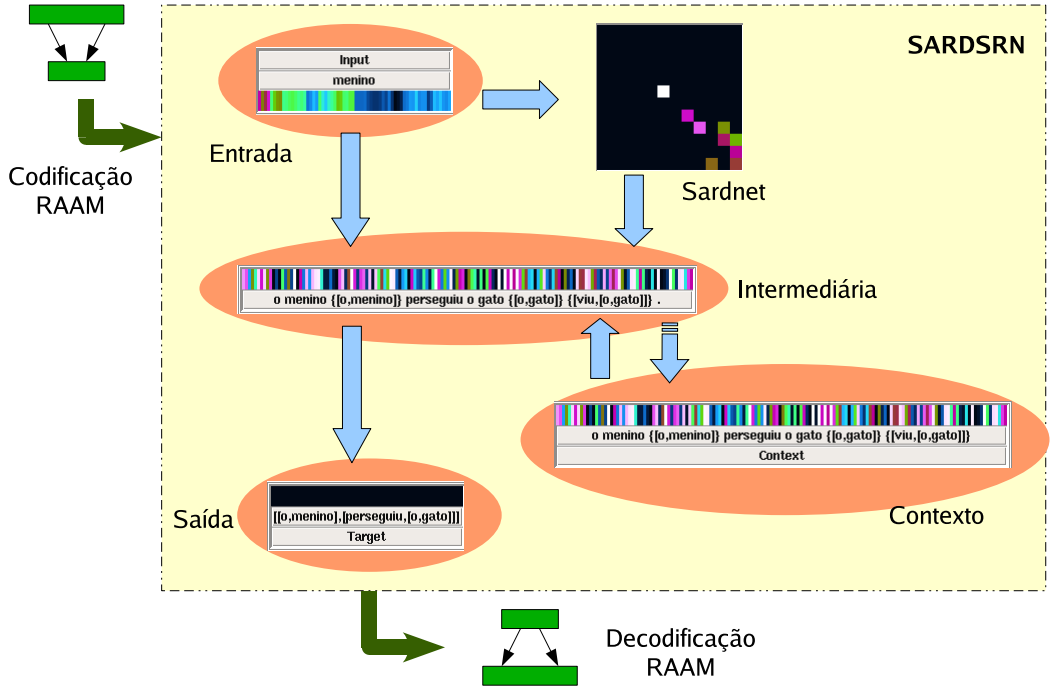

Figura 6.5: Componentes do SARDSRN.

e são armazenados em um arquivo (chamado speclex). As regras da gramática são representadas através de um arquivo (specraam) onde as relações básicas entre os componentes da gramática são apresentadas. Além de fornecer os valores dos terminais, o usuário necessita editar um arquivo contendo as cadeias para treinamento das frases. Após o treinamento, novos termos codificados são acrescentados ao arquivo de léxico, juntamente com os terminais. Esta codificação será utilizada pelo SARDSRN como entrada do sistema.

Após o treinamento, a saída do SARDSRN deve ser decodificada pela RAAM para a obtenção dos terminais da gramática. Estes são apresentados textualmente como resultado do reconhecimento.

6.2.3 Processamento prosódico-semântico com mapas auto-organizáveis

Segundo (LEWANDOWSKA-TOMASZCZYK, 1996), os aspectos fonéticos e semânticos são ligados por um sistema de representação de esquemas. Dentro de determinados contextos ocorre um agrupamento de palavras com um sentido criado por expectativas semânticas específicas (LEWANDOWSKA-TOMASZCZYK, 1996). A este fenômeno de composição do significado em torno de grupos de palavras faladas dá-se o nome de *semântica prosódica*.

Em paralelo a este conceito também deve ser colocado o de *prosódia afetiva*, que é a expressão não-verbal de um sentimento que está sendo comunicado com a fala (WAMBACQA; JERGER, 2004). Segundo (WAMBACQA; JERGER, 2004), a prosódia freqüentemente carrega o significado semântico de uma expressão de fala, o que vem a reforçar a idéia da semântica prosódica.

O estudo da prosódia pode auxiliar inclusive na análise de discurso. Trabalhos como o de (WARD, 2004), que examina as características prosódicas (seção 4.4.1) na análise pragmática, ou o de (BRAGA; MARQUES, 2004), que apresenta as características da pragmática da prosódia no discurso político. Estes estudos vêm somar-se aos conceitos

de semântica prosódica e prosódia afetiva no sentido de associar a análise prosódica à determinação de contextos de linguagem.

Um sistema desenvolvido por (KURIMO, 2002) usa redes SOM para indexação de documentos falados (notícias de rádio e televisão) usando métricas de análise semântica, semelhantes às usadas na análise pragmática. Este tipo de análise leva em conta a distribuição das palavras no documento transcrito por reconhecedor de voz.

Deve-se ressaltar ainda que o conceito de semântica aqui utilizado não se refere ao conjunto de conceitos desenvolvido ao longo do aprendizado da pessoa, mas apenas à característica formada pela análise da organização de uma frase. Este tipo de abordagem é citado por (ERDOGAN et al., 2005) como análise semântica rasa (*shallow semantic analysis*), semelhante aos sistemas de análise de segmentos apresentados na seção 2.3. Os sistemas que utilizam uma análise semântica completa são baseados em hierarquias de conceitos e ontologias, como os descritos em (ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006) e (RAYMOND et al., 2006).

A presente Tese sustenta que o processamento da análise semântica (rasa) deva ser influenciado pela prosódia, de forma a auxiliar na definição de contextos de compreensão. Com base nesta afirmação, propõe-se aqui o conceito de *prosódica-semântica*, que se define como sendo a influência da expressão da fala sobre a contextualização da linguagem. Com este conceito, tem-se a intenção de promover a investigação do conteúdo semântico embutido na prosódia. Aqui não é, portanto, utilizado apenas o conceito de vizinhança de palavras usado para a definição de contexto de uso, mas sim das características de similaridade da prosódia das palavras faladas que podem auxiliar na definição de contextos.

O subsistema prosódico-semântico aqui apresentado compõe-se basicamente de redes SOM para mapeamento das características prosódicas e das características fonéticas do sinal das palavras. A partir destes mapeamentos, é realizado um novo mapeamento, agora das relações semânticas das frases. Este mapeamento de frases irá auxiliar na decisão das frases reconhecidas no subsistema sintático, apresentado na seção 6.2.2.

As redes SOM foram escolhidas por sua característica de agrupamento por distância vetorial (veja seção 5.4.2). Os sistemas citados de análise semântica, tanto rasa (KURIMO, 2002; PARGELLIS et al., 2004; ERDOGAN et al., 2005) como completa (RAYMOND et al., 2006; ZHANG; HASEGAWA-JOHNSON; LEVINSON, 2006), utilizam basicamente métricas de proximidade vetorial para aproximação de conceitos por similaridade.

Todos os mapas aqui apresentados usam redes SOM de dimensão 10x10 (100 neurônios) seguindo o mesmo algoritmo, descrito na seção 5.4.2, para treinamento e reconhecimento. As implementações estão em linguagem C, mas o algoritmo de treinamento teve sua versão paralelizada para uso em processador dual e descrita em (MÜLLER; NAVAUX, 2005a).

A implementação paralela consistiu em desenvolver ramificações no algoritmo SOM, de forma a dividir os neurônios em dois grupos cada vez que ocorrer laços de varredura. Para tanto, antes de iniciar um laço do programa, a metade do mapa é processada na segunda unidade de processamento e o restante fica na unidade raiz. Ao final do laço o processamento continua no nodo raiz. Um trecho típico é apresentado a seguir:

```
if (posicao == 1)
 {
  MPI_Send(&dt, 1, MPI_INT, 0, 1,MPI_COMM_WORLD);
  i = dt+1;
```

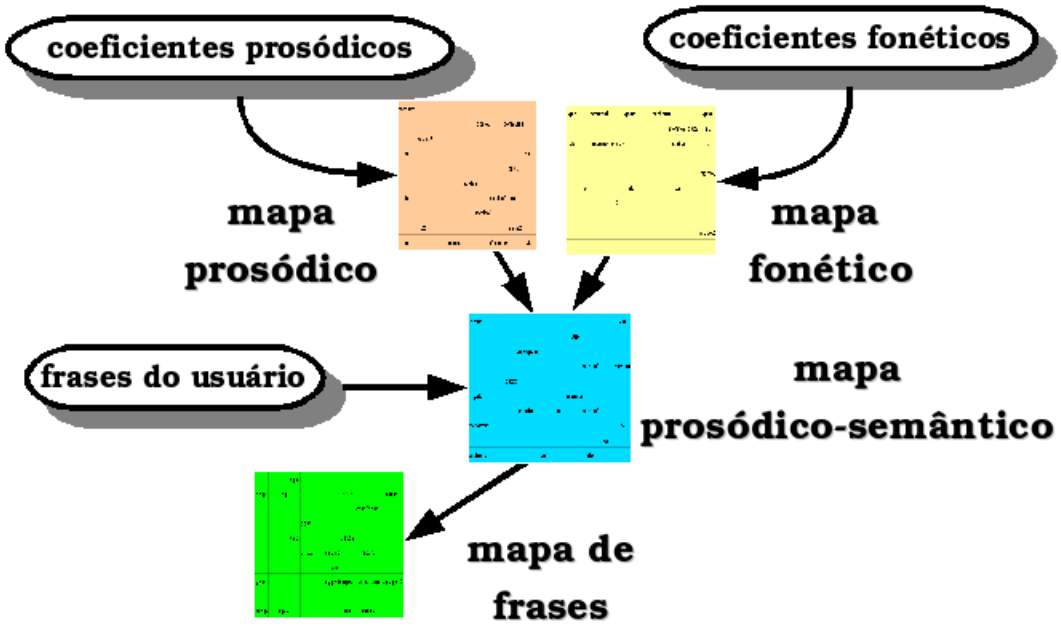

Figura 6.6: Organização dos mapas para agrupamentos lingüísticos, prosódicos e de frases.

```
  dt = dimt;
 }
if (posicao == 0)
 {
  result = MPI_Recv(&dt, 1, MPI_INT, 1, 1,
    MPI_COMM_WORLD, &status);
  if( result != MPI_SUCCESS )
    indiverr( "communication error" );
 }
for(i = 0; i < dt; i++) { ... }
```

Na primeira rede obtém-se um mapa das relações prosódicas das palavras envolvidas, e na segunda, outro mapa das relações das frases. Estas informações (saídas das redes) são repassadas para os mapas respectivos de palavras e frases, a partir dos quais serão mapeadas as relações semânticas. Como mostra a figura 6.6, cada mapa recebe os coeficientes que indicam as variações da onda fundamental, extraídos da transformada ondeletas conforme descrito na seção 6.2.1.

O posterior reconhecimento é feito a partir do último mapa, referente às relações semânticas de frases, o qual indica qual frase-padrão (com um significado definido) é semelhante à frase a reconhecer. Com este mecanismo básico é possível realizar-se a identificação dos contextos aos quais as frases apresentadas se referem.

6.2.3.1 Mapa prosódico

O mapa prosódico recebe os sinais derivados da análise da onda fundamental da voz, conforme descrito na seção 6.2.1.2. A forma pela qual a fala é expressada provoca variações na onda fundamental. Estas variações dão informações sobre as pausas e acentos da linguagem.

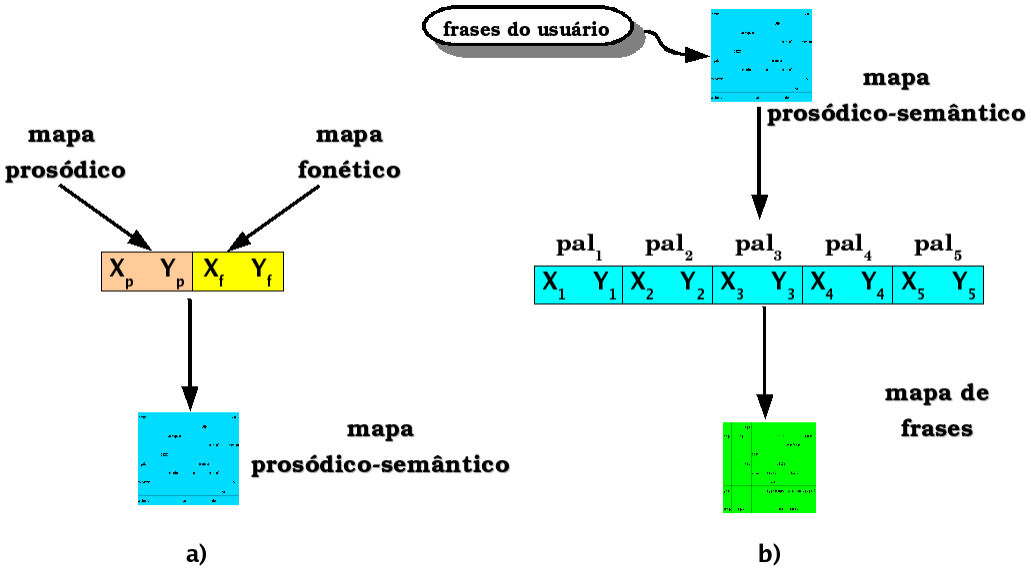

Figura 6.7: Composição das entradas do a) mapa semântico e b) mapa de frases.

O objetivo do mapa está na identificação e organização destas variações do pitch para as palavras faladas. O resultado do mapa é utilizado na composição da entrada de codificação das frases, juntamente com o resultante do mapa semântico.

6.2.3.2 Mapa fonético

Em paralelo ao processamento do mapa prosódico, o mapa fonético de palavras associa os dados provenientes dos coeficientes com informações extraídas através da transformada ondeletas, como descrito na seção 6.2.1.2. Este mapa organiza grupos de palavras segundo sua estrutura fonética, conforme a representação obtida a partir dos coeficientes das ondeletas.

Como resultado de seu processamento, são formados grupos de palavras com a constituição sonora similar. A organização destes grupos será utilizada na junção com as características prosódicas no mapa semântico.

6.2.3.3 Mapa prosódico-semântico

Para que cada palavra da frase possua uma significação da semântica prosódica, é proposto um mapa que reúna as ativações de determinada palavra para o mapa prosódico e para o fonético. Assim, o mapa prosódico-semântico une as características dos agrupamentos anteriores formando mais um nível de organização e possibilitando a distinção das palavras com propriedades prosódico-fonéticas semelhantes.

A entrada deste mapa possui 4 posições, compostas pela saída (posição no mapa) dos mapas prosódico e fonético para cada palavra verificada, como mostra a figura 6.7 a). Como uma palavra ativa um neurônio no mapa prosódico na posição (X_p, Y_p) e outro no fonético em (X_f, Y_f) , as coordenadas destes neurônios são compostas por quatro valores, os quais serão utilizados como padrão de entrada no mapa prosódico-semântico. O resultado da ativação de cada palavra neste mapa será utilizado no último mapa como entrada de cada palavra da frase selecionada.

6.2.3.4 Mapa de Frases

O mapa semântico de frases utiliza a indicação dos neurônios ativados no mapa prosódico-semântico a partir da frase indicada pelo usuário. O resultado deste mapeamento é a realização de agrupamentos segundo o conjunto de palavras mais prováveis dentro de um contexto de expressão.

Como entrada, este mapa recebe dois valores, que correspondem à posição no mapa semântico de cada palavra contida na frase fornecida pelo usuário (veja figura 6.7b). Para a atual implementação fixou-se um máximo de 5 palavras por frases, o que equivale a 10 entradas no mapa. O resultado deste mapa indica o contexto da frase indicada pelo usuário.

6.2.4 Módulo de Ponderações

O último módulo do processo possui uma função de avaliação das saídas do módulos sintático e prosódico-semântico. No módulo sintático, o sistema SARDSRN-RAAM retorna ao usuário uma taxa de erro de reconhecimento. No módulo prosódico-semântico, o mapa de frases resulta na ativação de determinado neurônio, que é comparado pelo módulo de ponderações com os neurônios dos padrões treinados. A posição da frase-padrão treinada mais próxima é retornada pelo módulo de ponderações, indicando o contexto semântico do reconhecimento.

Através deste módulo de ponderações, portanto, para dada frase apresentada ao sistema, a taxa de erro resultante da saída do SARDSRN-RAAM é comparada com o padrão de contexto mais próximo do mapa de frases. Por um lado, tem-se a validade da construção sintática da frase falada e, por outro, o contexto mais provável.

Com base nas informações obtidas, é retornado ao usuário a indicação de aceitabilidade ou não da frase falada. Trabalha-se com as seguintes hipóteses de resultados:

- uma taxa de erro sintático (indicada pelo SARDSRN-RAAM) acima de 0,5 indica rejeição da estrutura gramatical;
- uma distância maior que 2 entre a frase apresentada e as frases padrão no mapa de frases sinaliza rejeição de contexto;
- caso o módulo sintático e o módulo prosódico-semântico acusem rejeição, a frase será apontada como incompreensível;
- se não houver rejeição sintática, mas houver prosódico-semântica, haverá uma rejeição de contexto, indicando que a estrutura está correta;
- caso ocorra rejeição sintática, mas aceitação prosódico-semântica, a frase treinada mais próxima será indicada como a melhor escolha;
- não havendo rejeições, a frase reconhecida é apresentada.

Após a apresentação do resultado, se as ponderações indicarem algum tipo de rejeição parcial, cabe ao usuário tomar a última decisão. Nos demais casos, o nível de certeza permite ao sistema indicar a rejeição ou a aprovação.

A implementação apresentada neste capítulo foi objeto de simulações para demonstração de sua funcionalidade. No próximo capítulo serão apresentados os resultados do processamento dos módulos implementados.

7 SIMULAÇÕES DA IMPLEMENTAÇÃO DO *COMFALA*

A descrição das simulações irá seguir a mesma seqüência que a apresentada no modelo neurocognitivo (MNPAF). Num primeiro momento é descrito o histórico de simulações do processamento do sinal através das ondeletas. No seguimento é apresentado o treinamento reconhecimento dos coeficientes fonéticos no sistema SARDSRN-RAAM para realização do processamento sintático. A simulação da análise prosódico-semântica é apresentada através dos resultados dos mapas auto-organizáveis. Por fim é mostrado como a avaliação dos resultados de saída auxilia na determinação das frases e contextos de fala. Para análise dos resultados obtidos foram desenvolvidas técnicas para avaliação da organização proporcionada pelo processamento do sinal e pelo treinamento das redes neurais.

Para proceder os testes apresentados neste capítulo, foi selecionado um locutor para proceder a leitura de cinco frases de cinco palavras cada. As frases foram gravadas duas vezes, com expressões diferentes, sendo apenas as de primeira expressão usadas para treinamento.

Foram selecionadas treze palavras (*a, o, da, do, que, viu, gato, gostou, mordeu, menina, menino, cachorro, perseguiu*) seguindo uma tradução das utilizadas em (MAYBERRY III; MIIKKULAINEN, 1999) para compatibilização com o sistema de processamento sintático. As palavras foram gravadas com microfone digital, encontrado na maioria dos computadores, e amostradas a 22 KHz. Elas não foram gravadas separadamente, mas de forma contínua na frase. Após as gravações, as palavras foram segmentadas manualmente.

O reconhecimento da fala de diversos locutores e a segmentação automática das palavras são alguns dos temas de pesquisa mais relevantes nesta área. Apesar disso, estes temas exigem um aprofundamento específico que foge ao escopo desta Tese. As simulações apresentadas a seguir limitam-se a demonstrar a plausibilidade do modelo COMFALA como sendo uma visão computacional do MNPAF acrescido de uma etapa de análise prosódica.

Todas as simulações aqui apresentadas foram realizadas num mesmo computador monoprocessado de 1GHz, com exceção dos mapas fonético e prosódico, que foram treinados em processador dual (2x1,3GHz). O tempo médio de obtenção dos coeficientes fonéticos foi de 1,116s e prosódicos de 0,210s para cada palavra. O sistema SARDSRN-RAAM teve o treinamento de sua rede RAAM concluído em 30 min e do SARDSRN em 1 hora. O reconhecimento por palavra neste sistema ficou em média de 0,123s. O treinamento nos mapas ficou em média 30 min e o reconhecimento com uma 0,064 s para cada um.

Deve ser observado que o desempenho do sistema de extração de ondeletas é implementado em Matlab e o SARDSRN-RAAM é associado a uma interface gráfica. Ambos

os sistemas podem ter seu tempo de execução reduzido caso sejam realizadas implementações próprias e sem interface gráfica e em linguagem C.

7.1 Metodologia de análise

As simulações realizadas neste capítulo seguem metodologias de análise criadas especificamente para este trabalho. A análise dos resultados do processamento do sinal é feita com base nos próprios mapas formados no processamento prosódico-semântico. Os resultados dos processamentos sintático e semântico tem por referência a análise de uma gramática criada para validação da resposta dos módulos implementados.

7.1.1 Análise do processamento do sinal

A eficiência da codificação realizada pela transformada ondeletas pode ser constatada através da coerência de representações similares entre padrões sonoramente semelhantes. Partindo do princípio que os sons semelhantes devem gerar coeficientes próximos, o cálculo de distância vetorial aplicado aos mapas auto-organizáveis utilizados no processamento da linguagem devem indicar a tendência à qualidade de representação.

Desta forma, arquivos de áudio codificados por ondeletas contendo as mesmas frases devem ativar o mesmo neurônio ou seus vizinhos numa rede SOM. Procedendo assim, acredita-se ser possível uma adequada avaliação da validade da representação do sinal.

Para efeito de análise, define-se que, dada a ativação de um neurônio no mapa pela representação de coeficientes fonéticos ou prosódicos, se a distância vetorial entre o neurônio do padrão de teste o padrão de referência treinado for maior que 2, a representação pode ser inadequada. Caso seja ativado o mesmo neurônio ou seus vizinhos de distância 1, indica-se que a representação é válida, correspondendo ao padrão de referência.

7.1.2 Análise dos processamentos da linguagem

Para realização das análises sintática e semântica, as palavras gravadas foram compostas em frases. Os coeficientes fonéticos e prosódicos das palavras são utilizados para a formação de frases. Para a organização das frases desenvolvidas, é proposta uma gramática cujas regras são:

$$S \leftarrow SN\ SV$$
$$SN \leftarrow art\ subs$$
$$SN \leftarrow prep\ subs$$
$$SV \leftarrow verbo\ SN$$

onde S é a sentença, SN é o sintagma nominal, SV o sintagma verbal, $subs$={*menino, menina, cachorro, gato*}, art={*o,a*}, $prep$={*do,da*}, $verbo$={*gostou, viu, perseguiu, mordeu*}. As frases elaboradas foram divididas em cinco grupos:

A - frases treinadas;

B - frases gramaticalmente corretas, não treinadas, e com SNs trocados antes e depois do SN;

C - frases não treinadas com a colocação do SV no final - nenhuma frase foi treinada para este caso;

D - frases com erros graves de estrutura, desfigurando a formação dos SNs e SV;

E - seqüências de palavras repetidas, sem a intenção de formar uma frase.

Para o grupo A, foram treinadas as frases:

1. o menino viu o gato;
2. o gato gostou da menina;
3. o cachorro mordeu o menino;
4. a menina gostou do cachorro;
5. o cachorro perseguiu o gato;
6. o gato viu o cachorro;
7. o menino mordeu a menina;
8. o cachorro mordeu o gato;
9. a menina perseguiu o menino.

O comportamento das análises será avaliado pela capacidade de reconhecimento de frases não treinadas. Para tanto, foram selecionadas frases que seguem a gramática apresentada, mas sem serem treinadas. Para cada verbo foram feitos exemplos de alteração que seguem a ordem: de SN antes do verbo, SN depois do verbo e SN antes e depois do verbo, repetidas para cada verbo. Por exemplo, para a frase *o menino viu o gato*, foram criadas variações mantendo só o primeiro SN (*o menino viu*...), apenas o segundo SN (...*viu o gato*) e unicamente o SV (...*viu*...). As frases do grupo B são:

1. o menino viu a menina;
2. a menina viu o gato;
3. o cachorro viu o menino;
4. o cachorro gostou da menina;
5. a menina gostou do menino;
6. o menino gostou do gato;
7. o gato mordeu o menino;
8. o cachorro mordeu o gato;
9. a menina mordeu o gato;
10. o menino perseguiu o gato;
11. o cachorro perseguiu a menina;
12. o gato perseguiu o cachorro.

Foram desenvolvidas para o grupo C algumas frases colocando-se o verbo no final, o que confronta a gramática proposta, uma vez que ela não possui verbos como terminais:

1. o cachorro do menino mordeu;
2. o gato da menina gostou;
3. o menino o gato viu.

Dado que a língua oral nem sempre é gramatical, foram elaboradas para o grupo D algumas frases com o verbo no início e alterações na organização dos artigos e substantivos:

1. gato o menina da gostou;
2. viu o menino o gato;
3. mordeu o o cachorro menino;
4. cachorro gato o perseguiu o.

Por fim, segue o grupo E com *frases* totalmente degeneradas:

1. o o o o o;
2. gato gato gato gato gato;
3. gostou gostou a menino menino.

Através do reconhecimento dos cinco grupos de frases apresentadas, espera-se que, naturalmente, ocorra uma progressiva degradação nas taxas de acerto, na mesma proporção das distorções elaboradas. Os grupos de frases servem, portanto, de instrumento de avaliação da capacidade de reconhecimento dos sistemas de análise sintática e semântica.

7.2 Extração das características da fala por ondeletas

A simulação aqui descrita tem dois grandes objetivos: extrair coeficientes fonéticos, que permitam uma caracterização fonema-grafema (associação som-escrita), e coeficientes prosódicos, para estimação do pitch e, por conseqüência, das características da prosódia inerente ao trecho de fala em análise. Para o processamento das ondeletas, foi utilizado um ambiente que disponibilizasse as bibliotecas de funções necessárias à execução e análise dos dados. Dados estes fatores, foi escolhido o ambiente Matlab para realização de todos os testes relativos a ondeletas.

7.2.1 Escolha do modelo de ondeletas

O primeiro passo para utilização das ondeletas foi a escolha de qual modelo de coeficientes para filtragem pela ondeleta-mãe. Após a avaliação de todos os modelos de ondeletas disponíveis no Matlab, em todas as variações de número de coeficientes, foi optado o Daubechies com 8 coeficientes, conhecido no ambiente como *db4* (veja figura 7.1). Aparentemente o valor 4 do nome está relacionado à quantidade de coeficientes positivos do filtro.

O critério de escolha do modelo de filtro foi pelo que mantivesse as características de forma e capacidade de audição do sinal após a aplicação da transformada. O *db4* foi o filtro que melhor se adaptou a este critério, uma vez que os demais causavam distorções na onda e/ou na capacidade de audição.

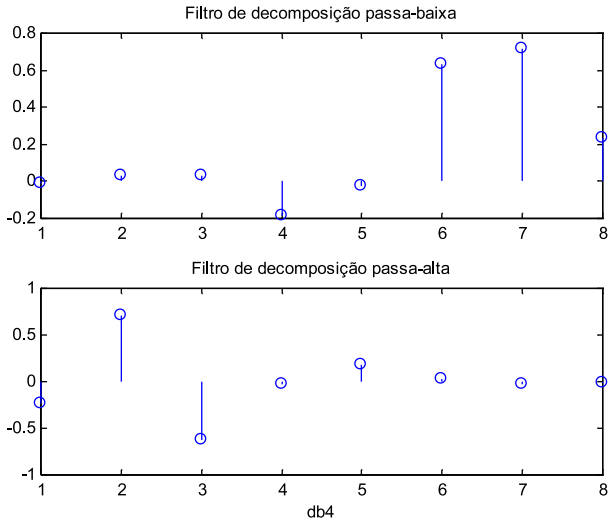

Figura 7.1: Filtros de decomposição Daubechies.

7.2.2 Obtenção dos coeficientes fonéticos

Numa primeira análise, a simples aplicação de uma transformada de ondeletas aparentemente bastaria para realizar a representação do sinal. Pegando-se os coeficientes de aproximação gerados sucessivamente por 3 níveis de decomposição, representariam uma faixa de sinal de cerca de 2700 Hz, o que permite a representação da onda fundamental da fala e suas principais formantes (veja figura 7.2).

Por outro lado, o número de coeficientes de ondeletas gerados nestes parâmetros é elevado, com uma média de 120 valores por palavra. O grande número de coeficientes também acarreta um alto custo computacional nas demais fases, necessitando processamento intensivo para treinamento das redes neurais.

Outro inconveniente que a simples decomposição acarreta é a grande variabilidade do número de coeficientes gerados, mesmo para palavras iguais, mas faladas de formas distintas ou por pessoas diferentes. A dificuldade de normalização do número de coeficientes, mesmo que atrelado a algum parâmetro como o número de sílabas, levou à busca da otimização de sua representação.

Nesta busca foram descartadas hipóteses como o uso do algoritmo MFCC com ondeletas pelos fatores expostos na seção 6.2.1.1. Por outro lado, foram encontradas alternativas válidas em (KIM; YOUN; LEE, 2000; FAROOQ; DATTA, 2004; RICOTTI, 2005), principalmente por afinarem com a descrição embasada biologicamente e matematicamente por (YANG; WANG; SHAMMA, 1992).

Conforme descrita na seção 6.2.1.1, a densidade espectral das folhas de uma árvore de decomposição de ondeletas pode ser utilizada para representação do sinal de fala. Esta abordagem resolve os problemas anteriormente elencados. O número de sub-bandas gerado pela árvore é pequeno, de apenas 8 ao terceiro nível de decomposição, equivalente ao número de folhas da árvore. Isso também resolve o problema da variabilidade do número de coeficientes, que sempre terá igual número ao das folhas da árvore, não importando o tamanho do sinal em análise.

A partir das constatações descritas, foi aplicada a extração de coeficientes de ondeletas com uso da densidade espectral normalizada das folhas da árvore de decomposição.

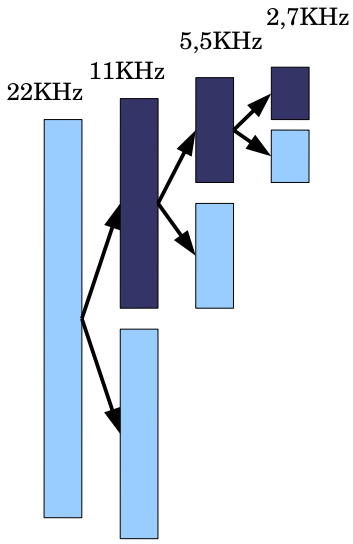

Figura 7.2: Amostragem dos coeficientes de aproximação em três níveis de decomposição.

Foram realizadas decomposições até o nível 6 para obtenção de 64 coeficientes, compatibilizando assim com o número de entradas das redes neurais de análise da linguagem.

A figura 7.3 apresenta o sinal original e coeficientes para as palavras *gato* (a) e *perseguiu* (b). O código Matlab utilizado na obtenção dos coeficientes fonéticos encontra-se no Anexo A.

7.2.3 Obtenção dos coeficientes prosódicos

Ao contrário dos coeficientes fonéticos, os prosódicos exigem que sejam extraídos da simples decomposição de ondeletas. Neste caso, são feitas decomposições aos níveis 2 e 3, seguindo a técnica de Kadambe descrita na seção 6.2.1.2.

Após as decomposições, os pontos de máximo são obtidos através da comparação entre os coeficientes do nível 3 que são iguais ou maiores que 80% do tamanho de seus correspondentes no nível 2. Para adequação do tamanho dos vetores de comparação, é feita uma decimação diádica do nível 2, a mesma que é utilizada para a transformada de ondeletas.

Para exemplificar esta etapa, as figuras 7.4 e 7.5 mostram, respectivamente para as palavras *gato* e *perseguiu*, nos seus itens a), os pontos de máximo encontrados pela comparação dos níveis 2 e 3. Nos itens b) das figuras são apresentados os cálculos dos ciclos, mostrando os pontos de início e fim. Os itens c) mostram a estimação do pitch através da variação de freqüência (onda quadrada superior) frente aos ciclos encontrados. Nos itens d) são apresentados os coeficientes resultantes.

Convencionou-se também a obtenção de 64 valores para representação prosódica. Como a geração de coeficientes de ondeletas é variável, caso a quantidade gerada seja maior que 70% do previsto, a transformada ondeletas *haar* é aplicada sucessivamente até atingir o patamar desejado, em torno de 64 valores. Caso seja menor, é completado com zeros, caso ainda seja maior, o excesso é ignorado. A ondeleta *haar* foi aplicada por seus coeficientes formarem uma onda quadrada, adequada à suavização da onda de variação de freqüência (veja ondas de pitch nas figuras 7.4c e 7.5c). O código Matlab utilizado na extração dos coeficientes prosódicos encontra-se no Anexo B.

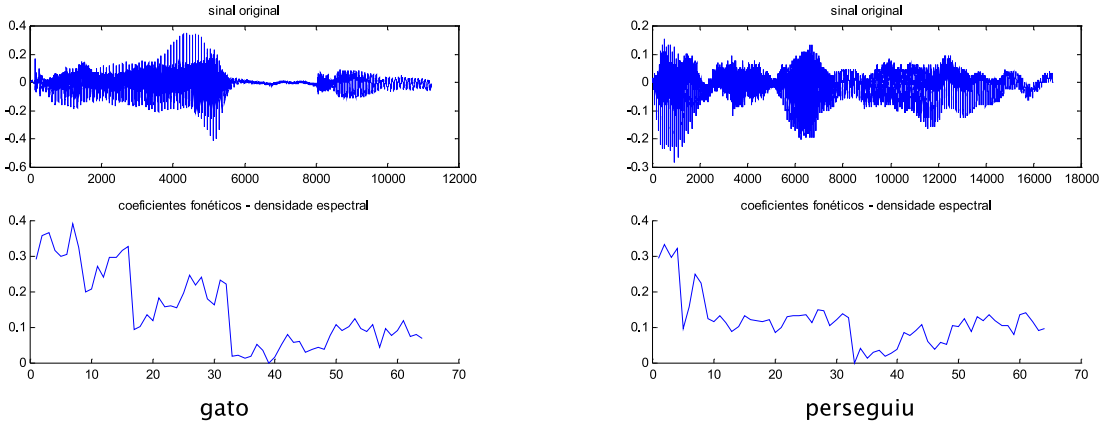

Figura 7.3: Geração de coeficientes fonéticos para a palavra *gato* em a) e *perseguiu* em b).

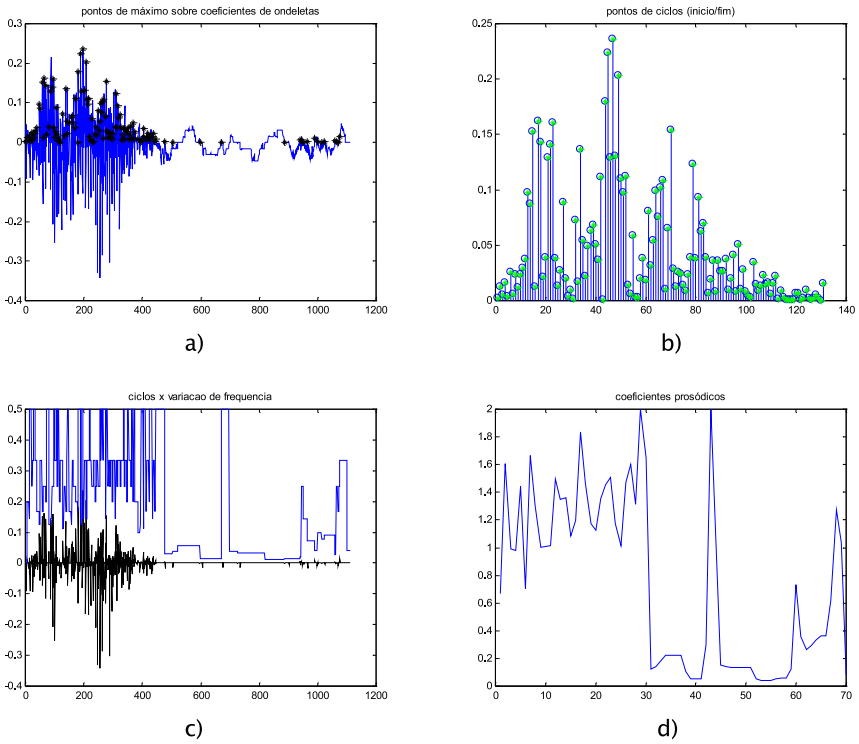

Figura 7.4: Passos de geração dos coeficientes prosódicos para a palavra *gato*: a) identificação dos pontos de máximo; b) cálculo dos ciclos entre os máximos; c) ciclos x variação da freqüência (onda superior); d) coeficientes obtidos.

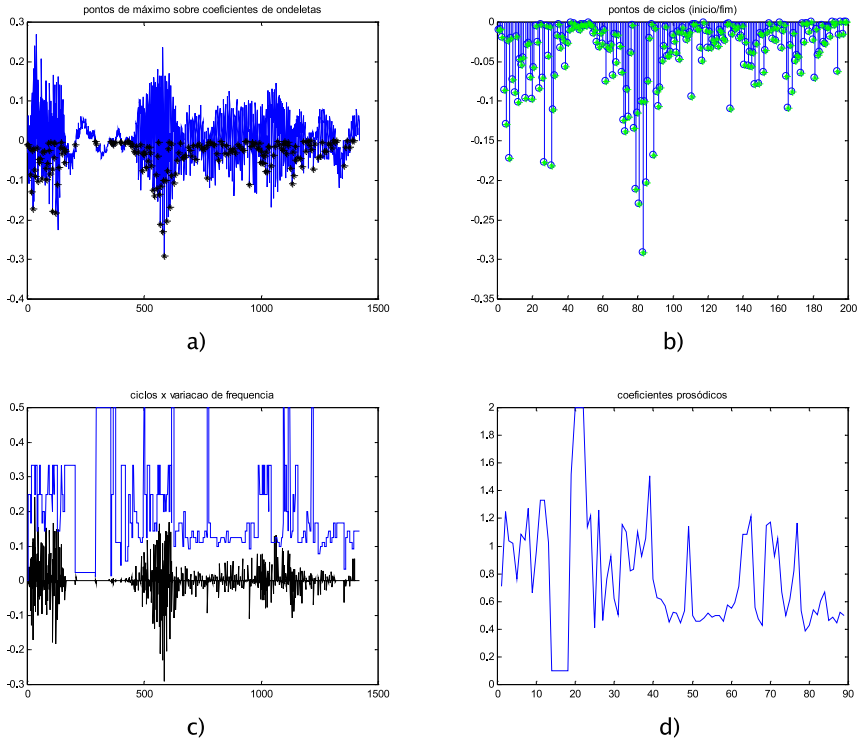

Figura 7.5: Geração dos coeficientes prosódicos para a palavra *perseguiu*, seguindo os mesmos passos da figura 7.4.

7.3 Análise sintática com o SARDSRN-RAAM

Os coeficientes fonéticos são aplicados diretamente como valores de representação das palavras do léxico do sistema SARSRN-RAAM. Seu uso inicia com a obtenção da codificação dos símbolos necessários à representação da gramática pela rede RAAM.

Cada símbolo terminal da gramática possui uma representação por coeficiente fonético e estão editados no arquivo *speclex*. As relações da gramática são compostas das relações possíveis entre os sintagmas e estão anotados no arquivo *specraam*. Quando a RAAM faz o treinamento do léxico, constrói novas entradas no léxico, colocando as codificações relativas às relações possíveis da gramática. No atual experimento, a rede RAAM foi treinada em 1.103 épocas, quando o erro de saída ficou estabilizado em 0.01. A figura 7.6 demonstra este processo de codificação RAAM.

Após a codificação pela RAAM, os valores contidos no arquivo *speclex* são usados como léxico de entrada no SARDSRN, onde são treinados em sua ordem temporal de uso nas frases. O arquivo de padrões de treinamento contém uma representação textual das frases a serem treinadas. Caso as frases de treinamento não sigam a gramática, não será possível obter a redução da taxa de erro de saída do sistema. O treinamento das frases foi procedido até o erro de saída chegar a 0,005, o que implicou em 1.002 épocas. A interface do SARDSRN-RAAM é mostrada na figura 7.7.

Quanto ao reconhecimento, foi procedida a metodologia de análise conforme os grupos de complexidade gramatical. Como esperado, não houve problemas quanto ao reconhecimento do grupo A, composto pelas frases treinadas. A taxa de erro resultante no SARDSRN-RAAM para todas as frases foi relativamente homogênea, ficando em 0,3. Este valor não é menor provavelmente devido a problemas de decodificação inerentes à

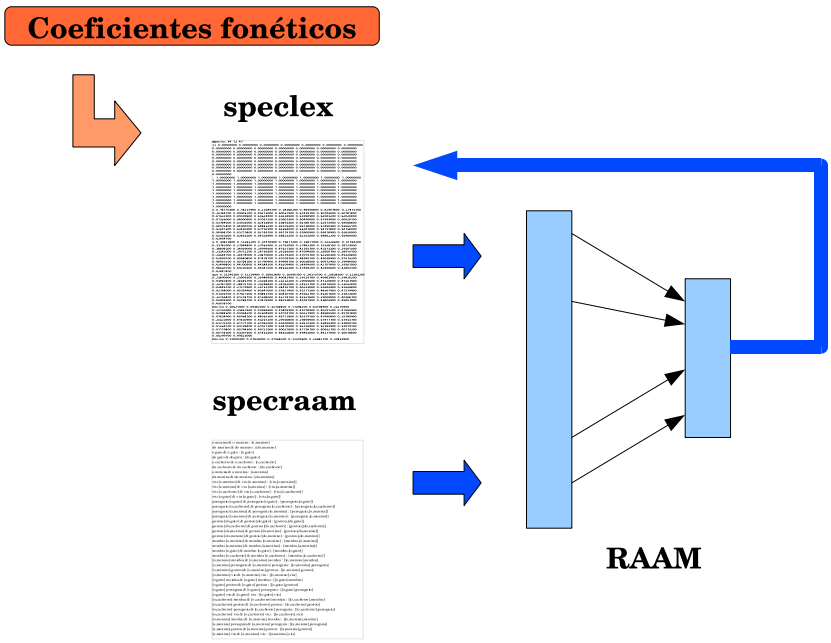

Figura 7.6: Codificação RAAM para treinamento no SARDSRN.

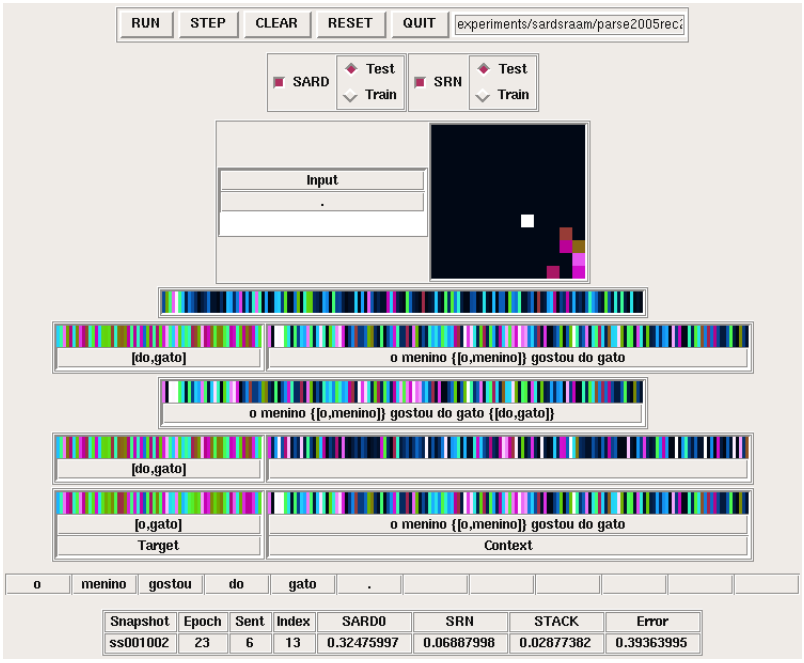

Figura 7.7: Interface do SARDSRN-RAAM.

rede RAAM, onde a transformada inversa proporcionada pela rede neural nem sempre garante encontrar nodos terminais, como já relatado por (LEVY; POLLACK, 2001).

O grupo B, por outro lado, apesar de não ser de frases treinadas, também obteve reconhecimento com taxa de erro 0,3, apresentada pelo SARDSRN-RAAM. Isso foi possível, uma vez que as construções S, SN e SV foram mantidas, embora fora da ordem previamente estabelecida. Como a codificação RAAM não foi afetada por ser mantida a estrutura gramatical, verifica-se então que o SARDSRN é robusto o bastante para realizar a aproximação de seqüências desconhecidas com as treinadas.

De fato, isso pode ser comprovado pelos exemplos demonstrados:

- a frase com a alteração do segundo SN *o menino viu a menina* teve como saída do SARDSRN a frase *o menino mordeu a menina* (figura 7.8a), aproximando assim uma frase desconhecida com aquela treinada com maior número de elementos similares;
- a frase com mudança do primeiro SN *a menina viu o gato*, que o sistema retorna como *a menina perseguiu o gato* (figura 7.8b), frase não treinada, no entanto foi obtida textualmente pela aproximação dos componentes mais prováveis;
- a frase com a manutenção do SV e alteração de ambos SN, *o cachorro viu o menino* foi aproximada para frase treinada *o cachorro mordeu o menino* (figura 7.8c)*;*
- outro caso interessante é a determinação exata da frase (relembrando, não treinada) *o cachorro gostou da menina*, que retornou este texto (figura 7.8d); o mesmo ocorreu com as frases *a menina gostou do menino* e *o gato mordeu o menino*.
- por outro lado, a frase *o menino gostou do gato* retornou como *o menino viu o gato*, que foi o comportamento geral das frases do grupo B, aproximando frases desconhecidas aos padrões treinados.

O grupo C, que tem frases com verbo no final indo contra a estrutura gramatical. Portanto, é natural a obtenção de erros 0,5 e 0,6 na saída do sistema. Com taxas acima de 0,5 o SARDSRN não consegue fornecer uma saída textual coerente para nenhuma das frases do grupo.

O reconhecimento para o grupo D também obteve elevados índices de erro. A única exceção é para a frase *viu o menino o gato*, que não afeta a construção nem do SN, nem do SV, mas do S, ocorrendo a inversão dos termos. Talvez seja esta a razão para seu reconhecimento, com taxa 0,3, identificando a frase *a menina perseguiu o gato* como sua equivalente. As demais frases não possuíram saídas de texto coerentes.

Como previsto, o grupo E não obteve acertos, uma vez que não possuem qualquer relação gramatical que sirva como referência significativa para o reconhecimento das frases. As taxas de erro ficaram em 0,6 e 0,5 e sem uma saída textual.

Os resultados do reconhecimento confirmam a robustez do sistema, como afirmado em (MAYBERRY III; MIIKKULAINEN, 1999). A única ressalva que deve ser feita ao SARDSRN-RAAM é quanto à necessidade de entrada textual para o treinamento da rede RAAM.

7.4 Análise prosódico-semântica com os mapas auto-organizáveis

Uma vez obtidos os coeficientes fonéticos e prosódicos, eles são utilizados como padrão de treinamento em dois mapas: um para agrupamento de características do que é

Figura 7.8: Diferentes frases reconhecidas no SARDSRN.

falado e outro de como é falado. Estes dois primeiros mapas necessitam utilizar a implementação paralela da rede SOM (veja seção 6.2.3), uma vez que possuem um elevado número de entradas (64), o que demanda mais processamento para convergência da rede. Isso porque o número de pesos por neurônio é igual ao das entradas, aumentando, assim, o tempo de adaptação da rede.

Os outros dois mapas, semântico e de frases, possuem poucas entradas, respectivamente 4 e 10, e portanto demandam menos tempo de treinamento. Estes mapas, então, não necessitam usar o SOM paralelo. Todos os mapas possuem dimensão 10x10 (100 neurônios) e foram treinados com raio igual a 4 e taxa de treinamento igual a 0,7.

Os mapas que serão analisados a seguir foram treinados seguindo a metodologia apresentada na seção 7.1. Assim, os coeficientes fonéticos e prosódicos das palavras da gramática foram treinados nos respectivos mapas. Os resultados dos mapas fonético e prosódico foram utilizados como entrada para o mapa semântico. Por fim, os resultados do mapa semântico foram utilizados para codificação das frases do grupo A, as quais formaram padrões para entrada no mapa de frases. Nas figuras dos mapas, os padrões treinados estão escritos em negrito e os reconhecidos estão em fonte normal. Os padrões reconhecidos com exatidão não estão representados, ou seja, só são apresentados em fonte normal os padrões que não obtiveram seu reconhecimento exato. Isso foi colocado como forma de análise das possíveis distorções de reconhecimento.

7.4.1 Treinamento e reconhecimento no mapa fonético

O mapa fonético foi treinado por um milhão e quinhentas mil épocas com algoritmo SOM paralelo, quando então os padrões estabilizaram nos mesmo neurônios vencedores. Para o reconhecimento foram apresentados, além dos padrões treinados, outros padrões com *a mesma palavra falada numa segunda gravação*.

As palavras contidas nos arquivos de áudio, tanto no treinamento como no reconhecimento, foram segmentadas à mão. Toda a segmentação possui uma margem de erro, que

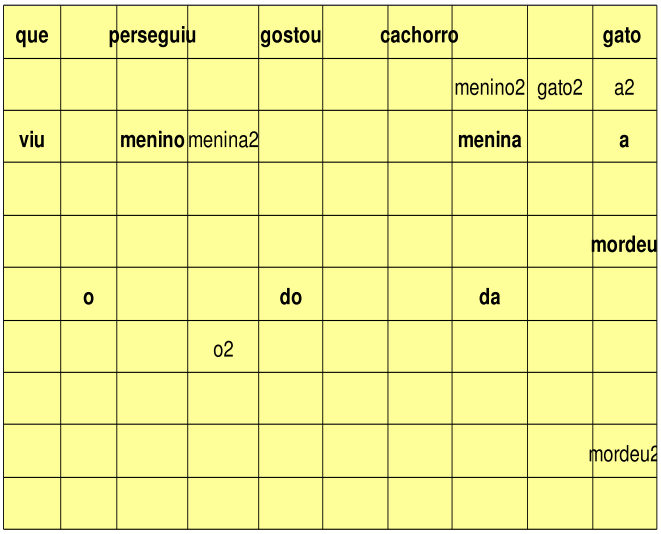

Figura 7.9: Mapa fonético.

pode afetar o reconhecimento da palavra, isso porque que na fala contínua ocorre muitas vezes a sobreposição de fonemas, impedindo uma adequada segmentação.

A figura 7.9 apresenta as relações entre as palavras treinadas (em negrito) e as reconhecidas. Os padrões que não estão repetidos indicam o reconhecimento exato no neurônio treinado. Esta relação também demonstra as propriedades da aplicação da transformada de ondeletas para extração das características do sinal. Isso porque a proximidade entre os padrões significa a semelhança entre os vetores de coeficientes obtidos.

Observa-se que numa primeira faixa (linha superior) foram organizados os padrões iniciados por *q*, *p*, *c* e *g*. Os padrões iniciados por *v*, *m* e *a* ficaram numa faixa abaixo e os padrões com *o* e *d* ficaram em outro grupo numa faixa mais abaixo. Ocorreu ainda um agrupamento no canto superior direito, provavelmente por falta de um melhor refinamento nos parâmetros da rede. Isso provocou um agrupamento também dos padrões reconhecidos da segunda gravação (*menino2*, *gato2* e *a2*).

Cabe salientar ainda a ocorrência de uma representação cruzada do reconhecimento de *menino* e *menina*, onde houve uma troca dos posicionamentos nos padrões reconhecidos. Isso pode ser devido ao nível de decimação (6) proporcionado pelas ondeletas no processo de extração dos coeficientes. A cada decimação ocorre um pequeno *aliasing* que dá uma distorção no sinal. Como as vogais finais das palavras são de alta freqüência, pode ter havido uma distorção em sua representação.

7.4.2 Treinamento e reconhecimento no mapa prosódico

O mapa prosódico foi treinado em um milhão de épocas com algoritmo SOM paralelo, quando então atingiu a convergência, estabilizando os padrões. O processo de seleção das palavras usadas foi o mesmo do mapa fonético, com a única diferença de ter sido aplicado o processo de extração dos coeficientes prosódicos.

O reconhecimento do mapa prosódico está apresentado na figura 7.10, que apresenta a similaridade entre os coeficientes obtidos pela estimação do pitch através das ondeletas. O mapa não apresenta agrupamentos significativos indicando classes de entonação das palavras. Por outro lado, houve a sobreposição do padrão *a* e *menino*, e novamente o reconhecimento cruzado de *menino* e *menina* (veja *menina2*).

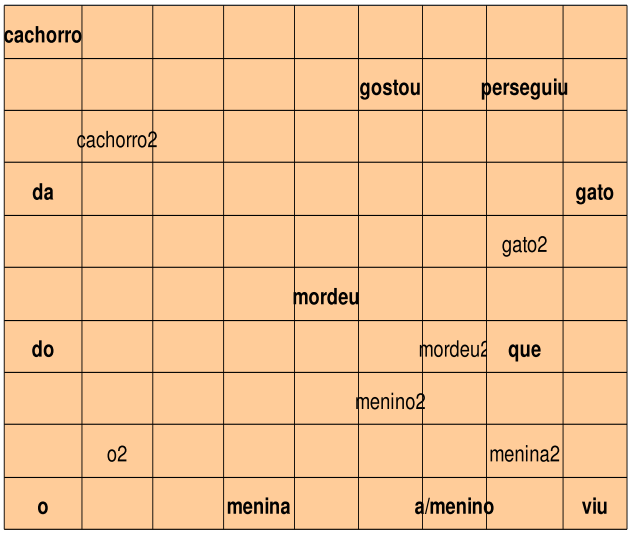

Figura 7.10: Mapa prosódico.

O fenômeno da sobreposição pode ter sido ocasionado pela dimensão inadequada do mapa, que necessitaria uma maior área, ou dos parâmetros de treinamento. O reconhecimento cruzado pode ser devido à semelhança entre as palavras.

Também observam-se distorções maiores no reconhecimento dos padrões de segunda gravação *cachorro2*, *mordeu2* e *menino2*, que podem ter sido causados pelo preenchimento de zeros na formação dos coeficientes prosódicos. Os zeros eram colocados quando os coeficientes não chegavam à quantidade estabelecida para o vetor de padrões (64).

Os resultados apresentados indicam que uma melhor análise pode ser realizada a nível de composições de grupos de palavras ou frases. Somente um futuro estudo da prosódia na fala contínua poderá esclarecer este ponto.

7.4.3 Treinamento e reconhecimento no mapa prosódico-semântico

O mapa semântico foi treinado com um milhão de épocas, concluindo com a estabilização dos pesos da rede. Seus padrões de treinamento são a composição das saídas dos mapas fonético e prosódico, como descrito na seção 6.2.3.3. Ele representa, portanto, a correlação entre as propriedades fonéticas e prosódicas identificadas anteriormente.

A figura 7.11 aponta algumas características importantes, como as palavras com um e dois fonemas agrupadas no canto inferior direito e no canto superior direito. Observa-se que as palavras de *segunda gravação* mantiveram o reconhecimento exato, ou numa distância vetorial menor ou igual a dois, como é mostrado nas palavras que possuem "2" ao final.

Pode-se inferir que a correlação fonético-prosódica auxilia na percepção de um contexto prosódico das palavras faladas. O mapa semântico, portanto, aparentemente mostra resultados que vem de encontro com o conceito de prosódica-semântica prosódica discutido na seção 6.2.3. As mesmas palavras faladas em diferentes frases em duas gravações (segunda gravação não treinada) com expressões diferentes podem corresponder a contextos prosódicos próximos.

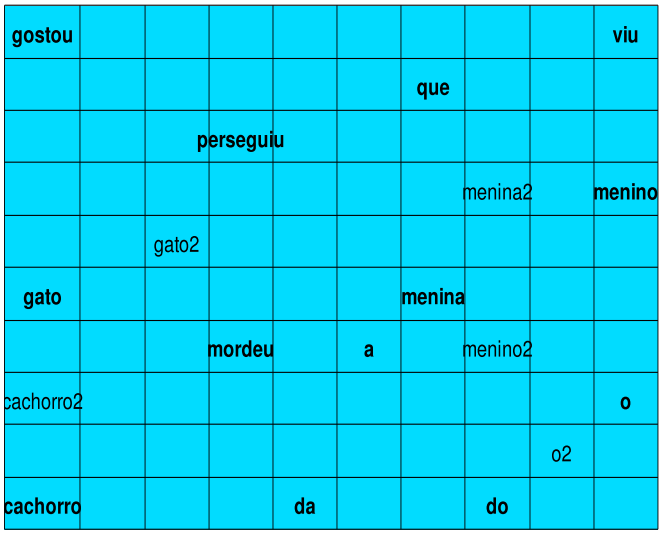

Figura 7.11: Mapa semântico.

7.4.4 Treinamento e reconhecimento no mapa de frases

O treinamento do mapa de frases se deu em um milhão de épocas, quando então convergiram os pesos. Os padrões apresentados à rede foram obtidos a partir da codificação das palavras treinadas no mapa semântico, ou seja, cada palavra da frase é um par ordenado que é composto num vetor para entrada na rede, como explicado na seção 6.2.3.4.

As frases apresentadas na figura 7.12 estão abreviadas, indicando substantivos e verbos. Por exemplo, a frase *o menino viu o gato* está abreviada como *m.v.g.* Conforme vem sendo adotado, os padrões treinados, no caso as frases do grupo A (frases treinadas), estão apresentadas em negrito.

O posicionamento das frases treinadas (grupo A) basicamente seguem as semelhanças de sua construção. Isso pode ser demonstrado por três padrões com componentes semelhantes: *o menino viu o gato* (m.v.g.), *o gato viu o cachorro* (g.v.c.) e *a menina gostou do cachorro* (m.g.c.). As palavras *menino/menina* e *cachorro* auxiliam no agrupamento observado no canto inferior esquerdo do mapa.

O reconhecimento do grupo B, com frases gramaticalmente corretas comportam-se como comentado para o grupo A, sendo aproximado dos padrões treinados pelos componentes semelhantes. Por exemplo, as frases *o cachorro viu o menino* (c.v.m.) e *o gato mordeu o menino* (g.m.m.) foram reconhecido como próximos ao padrão *o cachorro mordeu o menino* (c.m.m). As relações entre *cachorro* e *menino* e o termo *mordeu o menino* foram fundamentais para a indicação desta proximidade, que não seria verificada por nenhum dos demais padrões treinados. O mesmo raciocínio pode ser aplicado às demais frases do bloco B, com exceção à frase *a menina gostou do gato* (m.g.g.) que não obteve proximidade que permitisse uma melhor definição.

O reconhecimento do grupo C, contendo frases com verbo no final, afora a frase *o gato da menina gostou* (g.m.g.2) que ficou próximo à frase *a menina perseguiu o menino* (m.p.m.), os demais não ficaram próximos aos padrões treinados. A proximidade verificada aqui pode indicar uma proximidade prosódica, sem apontar uma relação gramatical.

Pode-se inferir para o grupo D, que possui grandes alterações gramaticais, e para o grupo E, das palavras em seqüência, a mesma linha de análise realizada para o grupo C. Apenas relações prosódicas podem justificar a proximidade de frases como *viu o menino*

			c.g.m.						
c.p.g.		c.m.g.					c.p.m.		**c.m.m.**
								c.v.m./g.m.m.	
				g.g.m.					
			v.m.g.				g.m.g.2		
				g.m.g.		c.m.m.2		**m.p.m.**	
						m.g.m.			
g.v.c.						m.g.g./o5	g.p.c.	m.v.m./m.m.g.	c.g.p./g2m2
m.v.g.		**m.g.c.**					m.c.m.	**m.m.m.**	

Figura 7.12: Mapa de frases.

o gato (v.m.g.) de *o gato gostou da menina* (g.g.m.).

7.5 Análise de resultados - as ponderações

As ponderações são realizadas comparando-se as saídas do SARDSRN-RAAM e a resposta do mapa de frases. Caso a taxa de erro indicada pelo SARDSRN-RAAM seja maior que 0,5 ou a distância no mapa seja maior que 2, teremos a rejeição.

Os exemplos de rejeição são as frases dos blocos C, D e E, onde apenas uma frase (*o gato da menina gostou*) conseguiu uma proximidade semântica no mapa de frases. Mesmo com esta proximidade, não é possível que seja garantido o contexto, uma vez que a proximidade da referida frase foi com *a menina perseguiu o menino*.

Para os padrões do grupo B, por outro lado, algumas vezes a análise semântica mostra-se melhor, como no caso de *a menina viu o gato*, que aproxima adequadamente de *o menino viu o gato*, o que, no SARDSRN-RAAM retorna como *a menina perseguiu o gato*. Por vezes também obtém-se o reconhecimento sintaticamente mais correto, como na frase *o cachorro gostou da menina*, que encontra uma resposta exata, mas semanticamente não está próximo aos padrões treinados no mapa de frases.

Observando-se este comportamento, percebe-se que o reconhecimento sintático possui mais eficiência quando a seqüência da frases possui alta probabilidade. Isso pode ser demonstrado no exemplo em que a frase *o cachorro gostou da menina*, pôde ser identificada com exatidão, uma vez que foi treinada a frase *o gato gostou da menina*, e sendo que, no mapa fonético (fig. 7.9) comprova-se que há proximidade entre a representação fonética de *gato* e *cachorro* nos padrões utilizados. Aliado a isto, como a seqüência da frase era conhecida, o padrão mais próximo a identificar foi o apresentado, uma vez que a frase que se tentava reconhecer não fora treinada.

Por outro lado, o reconhecimento semântico obteve mais sucesso nas representações onde a seqüência não foi suficiente para determinação da frase provável. No caso da frase analisada *a menina viu o gato*, apesar da seqüência ser muito semelhante à *o menino viu o gato*, a distância entre a representação fonética de *menino* e *menina* não permitiu à análise sintática realizar uma aproximação satisfatória. Em outras palavras, a aproximação pro-

Tabela 7.1: Resultado das ponderações.

grupo	**rec. sintático**	**rec. semântico**
A	sucesso	sucesso
B	parcial	parcial
C/D/E	falha	falha

sódica dos contextos utilizados compensou a diferença da representação fonética. Como se pode perceber na figura 7.10, a distância entre *menino* e *menina* é a metade daquela encontrada no mapa fonético, o que permitiu o reconhecimento semântico da frase para o contexto correto.

Quanto aos padrões do grupo A, ambos os sistemas conseguem realizar o reconhecimento perfeitamente. Este comportamento já era esperado, uma vez que as redes neurais são treinadas até atingirem a adequada representação aos padrões de treinamento.

Uma síntese das ponderações obtidas é apresentada na tabela 7.1, que demonstra a necessidade de um sistema para reorganização de frases, caso estas estejam nos grupos C, D ou E. No caso do grupo B o reconhecimento com sucesso é obtido pela compensação dos casos. Onde é parcial pela via sintática pode ser finalizado pela análise semântica e vice-versa, como exemplificado. No grupo A todos os casos são reconhecidos por ambas as análises.

8 CONCLUSÕES

Neste capítulo são retomadas as questões de pesquisa, reafirmando a necessidade do desenvolvimento realizado. Também é dada uma visão geral sobre as futuras derivações com base no modelo desenvolvido, as contribuições científicas desta Tese e futuras aplicações dos sistemas derivados do COMFALA.

Voltando às hipóteses de pesquisa:

- É viável o desenvolvimento de um sistema integrado conexionista para compreensão de fala?

Constatou-se que, melhor que um sistema integrado, sistemas componentes que se complementam parecem ser uma melhor forma de pesquisa e desenvolvimento, dada a complexidade envolvida no processamento da linguagem falada. Por outro lado, os sistemas conexionistas ou estocásticos parecem ser mais adequados por tratarem de padrões numéricos que podem ser obtidos do sinal de fala.

- A codificação da fala por transformadas de Fourier pode ser substituída por outra metodologia para melhor caracterização da linguagem falada contínua?

Como analisado na seção 4.3.1, foram constatadas algumas vantagens do uso da transformada ondeletas (wavelets) frente à transformada Fourier. Uma melhor forma de análise temporal em função das freqüências e a simplificação do processo de obtenção dos coeficientes, não necessitando usar o algoritmo MFCC, foram fatores que influenciaram no uso da ondeletas como técnica de codificação do sinal de fala.

- A análise da onda fundamental pelos algoritmos tradicionais é eficiente e suficiente para a obtenção dos dados prosódicos?

A análise realizada atualmente é suficiente, mas foi encontrada uma alternativa de estimação de pitch através da transformada ondeletas. Isso permitiu uma simplificação do processo, de forma semelhante ao discutido na questão anterior.

- Modelos conexionistas podem ser utilizados como alternativa ao HMM para modelagem da linguagem?

Conforme demonstrado no decorrer da Tese, os modelos neurais SRN e SARDNET podem ser utilizados para armazenamento de seqüências temporais. Isso permite que sejam aplicados como alternativa aos modelo ocultos de Markov.

- Modelos conexionistas podem ser uma alternativa viável à árvore de análise da linguagem?

A capacidade da rede neural RAAM de permitir a representação de organizações hierárquicas faz dela uma boa alternativa às árvores de parser probabilístico. A vantagem do uso da rede neural é o treinamento através de exemplos, o que otimiza o processo de modelagem da linguagem.

- Os dados prosódicos podem ser inseridos diretamente nos processos de análise sintática e semântica?

Na seção 6.1.3, constatou-se que a prosódia não interfere no processamento sintático, mas afeta a análise semântica. Neste sentido, o COMFALA propõe o uso da prosódia em conjunto com a análise semântica. Na implementação do módulo prosódico-semântico constatou-se que a correlação das características de forma e entonação do sinal auxiliam na definição de contextos de uso das frases.

- Como ocorre o processo natural da compreensão da fala?

Foi apresentado no capítulo 3 não propriamente a compreensão da fala, mas a forma de processamento da audição de frases. Uma melhor análise da compreensão de fala deve levar em conta outros fatores de influência no processo, tais como os sentidos, o movimento e a visão.

Repete-se, então a hipótese que definiu o desenvolvimento desta Tese:

- *Deve ser possível a constituição de um modelo computacional que tenha por base pesquisas da compreensão do processamento natural da compreensão da fala. Este modelo deve permitir uma representação temporal do sinal da fala e de sua prosódia. Os dados obtidos na representação do sinal devem servir de base para as análises sintática e semântica da linguagem. Estas análises devem ser construídas automaticamente, a partir de exemplos de fala.*

A hipótese foi comprovada através da construção de um modelo, denominado COMFALA, bem como o desenvolvimento ou aperfeiçoamento de protótipos para demonstrar sua funcionalidade.

8.1 A definição de um modelo

A presente Tese procura indicar uma modelagem computacional que represente a organização do processamento cerebral da audição de frases. O modelo proposto, denominado COMFALA, foi projetado de forma a automatizar o processo de geração de padrões de fala e com eles possibilitar as análises sintática e semântica. Estas análises, ao contrário de alguns sistemas de linguagem natural, não realizam um processo seqüencial, mas um processamento independente. Isso foi motivado pelo modelo neurocognitivo estudado (MNPAF).

A vantagem verificada pelo processamento independente das análises sintática e semântica foi a criação de um sistema de compensação, onde pode ocorrer de a análise sintática não corresponder ao padrão pretendido, mas a análise semântica obter maior

grau de certeza, e vice-versa. Para um sistema de linguagem falada a capacidade de análises alternativas é fundamental, devido à grande variação do sinal de fala e da estrutura das frases.

Os resultados obtidos com a implementação do modelo não pretende ser uma conclusão, mas uma linha de investigação a ser seguida no futuro. Pode-se dividir esta linha em quatro, seguindo os módulos do COMFALA. O primeiro campo de pesquisa é referente ao aprofundamento da extração de características do sinal por ondeletas.

Os resultados apresentados mostram-se promissores, apenas com distorções localizadas, mas uma melhor acurácia na distinção da onda fundamental e suas formantes devem ser investigadas para melhor representação das características fonéticas. Quanto à prosódia, acredita-se que ainda muito poderá ser aperfeiçoado, com a construção de modelos matemáticos mais robustos que representem de forma mais adequada as novas descobertas oriundas do campo neurocognitivo.

O segundo campo de pesquisa é na direção da análise sintática subsimbólica, realizada com a implementação de redes neurais. Este tipo de abordagem necessita de testes exaustivos com grandes bases de dados, visando o desenvolvimento de modelos que atendam a necessidades reais do reconhecimento de fala.

A terceira via de pesquisa segue no caminho da análise prosódico-fonética, iniciada com a presente Tese. Percebe-se a limitação da implementação realizada frente à necessidade de representação de diferentes estruturas de linguagem. Apesar da modelagem fonético-prosódica atender ao modelo neurocognitivo, sua implementação deve possibilitar a representação de fala contínua para um grande vocabulário. Frente a este desafio, novas implementações devem ser geradas para atender a estas necessidades.

O quarto caminho de pesquisa é referente à criação de métodos que se fazem necessários às ponderações de análise de saída dos sistemas implementados seguindo o modelo COMFALA. Como foi apontado nos resultados, os erros de estrutura de frase são determinantes para o bom reconhecimento. Isso vem de encontro ao MNPAF, o qual indica que, se verificada uma frase com estrutura errada, trata-se logo de corrigir esta organização antes de realizar outras relações.

Desta forma, também devem ser criados métodos que permitam a reorganização das frases apresentadas ao sistema, de maneira a constituir ciclos de correção de estrutura. Com esta propriedade, o módulo de ponderações seria responsável por tentativas de correção da frase apresentada.

O trabalho desenvolvido nesta tese tem sido publicado em Workshops (MÜLLER; NAVAUX, 2004, 2005a,b) e foi aprovado para apresentação no AIA'2006 (MÜLLER; NAVAUX, 2006), PROPOR'2006 (MÜLLER; SIQUEIRA; NAVAUX, 2006a) e IJCNN 2006 (MÜLLER; SIQUEIRA; NAVAUX, 2006b).

8.2 Contribuição científica

Como fundamento do COMFALA foi descrito o MNPAF, desenvolvido por Angela Friederici. Através da investigação da modelagem neurocognitiva, foi possível a organização dos dados relativos à importância da prosódia para a audição de frases.

Decorrente desta organização das informações sobre o modelo neurocognitivo, decidiu-se incorporar os novos dados ao MNPAF sem, no entanto, existirem no modelo original. Este procedimento é entendido como uma colaboração deste trabalho aos pesquisadores das áreas de neurociências. Além destes, também pretendeu-se constituir uma nova concepção para os sistemas computacionais de compreensão de fala, com a criação de um

modelo computacional baseado em estudos neurocognitivos.

Apesar do estudo das transformadas ondeletas (wavelets) serem um campo consolidado, sua aplicação ainda é pouco difundida na Computação. Neste sentido, aqui foi demonstrada a aplicabilidade da representação do sinal de fala por ondeletas. Os coeficientes gerados foram efetivamente utilizados como padrões nas redes neurais. Os testes realizados demonstraram ser possível a adoção da transformada ondeletas como uma metodologia de codificação de fala a ser usada em sistemas de processamento da linguagem.

Nesta Tese foi ainda resgatada a pesquisa com sistemas de análise conexionista (por redes neurais), através da utilização de um sistema criado para análise da linguagem apenas textual. Foi demonstrado que a propriedade de representação numérica dos sistemas conexionistas adequam-se a uma representação automatizada provida pela codificação do sinal de fala. Esta nova abordagem indica a possibilidade de aproveitamento de outros sistemas conexionistas ou estocásticos, desenvolvidos para língua escrita, para processamento da linguagem falada. Acredita-se que, com pequenas adaptações, antigos sistemas de análise de texto possam ser futuramente aproveitados para análise de fala.

Os mapas de análise semântica foram resgatados do trabalho de mestrado deste autor (MÜLLER, 1996) e aperfeiçoados de forma a representarem características do sinal de fala. Neste sentido, como na análise sintática, um sistema criado para processamento de linguagem escrita pôde ser adaptado para representação das relações lingüísticas contidas na fala.

No decorrer do desenvolvimento do trabalho os mapas semânticos se mostraram muito úteis como ferramentas de análise dos dados extraídos pelas ondeletas. Quando os padrões não são distribuídos adequadamente no mapa auto-organizável, sabe-se que há um problema de representação. Esta característica permitiu correções na codificação do sinal de forma a satisfazer as relações fonéticas entre as palavras analisadas. Os mapas semânticos podem ser, portanto, adotados futuramente como uma metodologia de validação de diferentes técnicas de codificação do sinal de fala.

Quanto ao modelo COMFALA, a sua concepção não foi feita na pretensão de realizar uma definição de um sistema, mas sim criar uma arquitetura básica que possa ser implementada através de diversos sistemas. Neste sentido, esta Tese procura apontar as linhas para desenvolvimento dos sistemas componentes do COMFALA: processamento do sinal de fala, análise sintática, análise prosódico-semântica e ponderações das análises.

Para cada uma dessas linhas, podem ser desenvolvidos futuramente sistemas em escalas cada vez maiores de robustez e acurácia na compreensão da fala. O COMFALA é, portanto, um caminho para a composição de sistemas, e não para um sistema único.

8.3 Possíveis aplicações e futuros aperfeiçoamentos

A pesquisa fundamental que está em discussão nesta Tese refere-se à necessidade de desenvolvimento de interfaces de fala para sistemas computacionais. A importância desta investigação reside na facilidade de operação de sistemas que advirá com a adoção do diálogo falado.

Como é constatado nesta Tese, o diálogo falado com sistemas computacionais não é um desenvolvimento trivial, mas que necessita de diversas etapas de processamento. Neste sentido, a definição do COMFALA pretende ser um parâmetro para a construção de sistemas que atendam às necessidades do processamento da fala, ao menos na parte relativa à compreensão da linguagem falada. Para a implementação de sistemas de diálogo falado faltaria a definição de um modelo de geração de fala.

Dentro do escopo de um sistema de diálogo, o COMFALA não aborda a análise de discurso ou geração de fala. Entende-se aqui ser, neste momento, mais importante a definição de uma interface de entrada que permita o acionamento e interação mais amigável dos sistemas computacionais.

Para viabilização dos protótipos aqui apresentados, permitindo um uso real da concepção apresentada na forma do COMFALA, seriam necessários ainda diversos aperfeiçoamentos, como já indicados neste capítulo. Dentre os aperfeiçoamentos que permitiriam o uso cotidiano das idéias aqui contidas seria uma técnica de segmentação do sinal, de forma a permitir a obtenção dos segmentos de análise em tempo de execução.

Outro desenvolvimento necessário é o ajuste dos sistemas para um grande léxico, permitindo um vocabulário realista de uso. Para tanto, a representação por mapas aqui utilizada pode não ser a melhor hipótese de implementação, dado o custo computacional envolvido no treinamento do léxico.

Também outras formas de estruturação semântica devem ser propostas. O uso de ontologias para a definição de contextos de uso pode permitir uma melhor organização do conhecimento contido no sistema. A partir deste ponto poderia ser realizada a análise pragmática da fala.

Por fim, não se deseja que este seja um modelo estático, ao contrário. À medida que os estudos neurocognitivos explicitarem novas relações do processamento cerebral, o COMFALA deve ser revisto e ampliado para atender às novas descobertas e assim permitir sua evolução.

REFERÊNCIAS

ABRAHÃO, P. R. C.; LIMA, V. L. S. de. Um Estudo Preliminar de Metodologias de Análise Semântica da Linguagem Natural. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DE PORTUGUÊS ESCRITO E FALADO, 2., 1996, Curitiba. **Anais...** Curitiba: CEFET-PR, 1996. p.139–148.

AMARO JUNIOR, E.; YAMASHITA, H. Aspectos básicos de tomografia computadorizada e ressonância magnética. **Revista Brasileira de Psiquiatria**, [S.l.], v.23, p.2–3, mai. 2001.

ARAÚJO, D. B. de; CARNEIRO, A. A. O.; BAFFA, O. Localizando a atividade cerebral via magnetoencefalografia. **Ciência e Cultura**, São Paulo, v.56, p.38–40, jan.-mar. 2004.

AVCI, E.; TURKOGLU, I.; POYRAZ, M. Intelligent target recognition based on wavelet packet neural network. **Expert Systems with Applications**, [S.l.], v.29, p.175–182, 2005.

BARKER, K. **Semi-Automatic Recognition of Semantic Relationships in English Technical Texts**. 1998. PhD Thesis — University of Ottawa, Ottawa.

BILMES, J. A. **Natural Statistical Models for Automatic Speech Recognition**. 1999. PhD Thesis — University of Berkeley, Berkeley.

BORNKESSEL, I. D.; FIEBACH, C. J.; FRIEDERICI, A. D. On the cost of syntactic ambiguity in human language comprehension: an individual differences approach. **Cognitive Brain Research**, [S.l.], v.21, p.11–21, Sept. 2004.

BOSTANOV, O.; KOTCHOUBEY, B. Recognition of affective prosody: continuous wavelet measures of event-related brain potentials to emotional exclamations. **Psychophysiology**, [S.l.], v.41, p.259–268, 2003.

BOYE, J.; GUSTAFSON, J.; WIRÉN, M. Robust spoken language understanding in a computer game. **Speech Communication**, Amsterdam, v.48, p.335–353, Mar.-Apr. 2006.

BRAGA, D.; MARQUES, M. A. The Pragmatics of Prosodic Features in the Political Debate. In: SPEECH PROSODY, 2004. **Proceedings...** [S.l.: s.n.], 2004. Disponível em: <http://www.isca-speech.org/archive/sp2004>. Acesso em: ago. 2005.

BRUIN, J. de; PREEZ, J. du. Automatic language recognition based on discriminating features in pitch contours. In: IEEE SOUTH AFRICAN SYMPOSIUM SIGNAL PROCESSING, 1993. **Proceedings...** [S.l.: s.n.], 1993. p.133–138.

BUØ, F. **FeasPar - A Feature Structure Parser Learning to Parser Spontaneous Speech**. 1996. PhD Thesis — Universität Karlsruhe, Karlsruhe.

BUØ, F.; WAIBEL, A. Search in a Learnable Spoken Language Parser. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, ECAI, 14., 1999. **Proceedings...** [S.l.: s.n.], 1999. v.2, p.629–632.

CARBONEL, J. G.; HAYES, P. J. Robust Parsing Using Multiple Construction-Specific Strategies. In: BOLC, L. (Ed.). **Natural Language Parsing Systems**. Berlim: Springer-Verlag, 1987.

CARNERO, B.; DRYGAJLO, A. Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms. **IEEE Trans. Signal Processing**, [S.l.], v.47, p.1622–1635, June 1999.

CASAGRANDE, R. Redes Neurais do tipo TDNN. In: CABRAL JR., E. F. (Ed.). **Redes Neurais Artificiais**. São Paulo: Ed. dos autores, 1999.

CHEN, S.-H.; WANG, J.-F. Extraction of pitch information in noisy speech using wavelet transform with aliasing compensation. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, ICASSP, 2001. **Proceedings...** [S.l.: s.n.], 2001. v.1, p.89–92.

COVOLAN, R.; ARAÚJO, D. B. de; SANTOS, A. C. dos; CENDES, F. Ressonância magnética funcional: as funções do cerébro reveladas por spins nucleares. **Ciência e Cultura**, São Paulo, v.56, p.40–42, jan.-mar. 2004.

CUNHA, C. F. **Gramática da Língua Portuguesa**. Rio de Janeiro: Fename, 1982.

DAUBECHIES, I. **Ten lectures on wavelets**. [S.l.]: Siam, 1992.

ECKSTEIN, K.; FRIEDERICI, A. D. Late interaction of syntactic and prosodic processes in sentence comprehension as revealed by ERPs. **Cognitive Brain Research**, [S.l.], v.25, p.130–143, 2005.

ELMAN, J. Finding structure in time. **Cognitive Science**, [S.l.], v.14, p.179–211, 1990.

ERDOGAN, H.; SARIKAYA, R.; CHEN, S. F.; GAO, Y.; PICHENY, M. Using semantic analysis to improve speech recognition performance. **Computer Speech and Language**, London, v.19, p.321–343, 2005.

ESCA WORKSHOP ON DIALOGUE & PROSODY, 2002, Amsterdam. **Proceedings...** [S.l.: s.n.], 2002. v.36.

EVANGELISTA, G. Pitch-Synchronous Wavelet Representations of Speech and Music Signals. **IEEE Trans. Signal Processing**, [S.l.], v.41, p.3313–3330, 1993.

FAROOQ, O.; DATTA, S. Wavelet based robust sub-band features for phoneme recognition. **IEEE Trans. Vision, Image and Signal Processing**, [S.l.], v.151, p.187–193, June 2004.

FERREIRO, E. **Cultura escrita e educação**. Porto Alegre: Artmed, 2001.

FOZ, F. B.; SILVINO, A. P.; RONDÓ, A. G.; BURSZTYN, C. S.; RODELLA, E. C.; LUCCHINI, F. L. P.; FUINI, M. G. **Análise da atividade cerebral durante a compreensão de charadas**. Disponível em: <http://www.enscer.com.br/pesquisas/artigos/charadas/charadas.html>. Acesso em: ago. 2005.

FRANÇA, A. I. Um flagrante da linguagem no cérebro. **Ciência Hoje**, [S.l.], v.26, p.20–25, jan.-fev. 2005.

FRIEDERICI, A. D. The Time Course of Syntatic Activation during Language Processing: a model based on neuropsychological and neurophysiological data. **Brain and Language**, [S.l.], v.50, p.259–281, 1995.

FRIEDERICI, A. D. Towards a neural basis of auditory sentence processing. **Trends in Cognitive Sciences**, [S.l.], v.6, p.78–84, 2002.

FRIEDERICI, A. D. Processing local transitions versus long-distance syntactic hierarchies. **Trends in Cognitive Sciences**, [S.l.], v.8, p.245–247, 2004.

FRIEDERICI, A. D.; ALTER, K. Lateralization of auditory language functions: a dynamic dual pathway model. **Brain and Language**, [S.l.], v.89, p.267–276, 2004.

FRIEDERICI, A. D.; KOTZ, S. A. The brain basis of syntatic processes: functional imaging and lesion studies. **NeuroImage**, [S.l.], v.20, p.S8–S17, 2003.

FRIEDRICH, C. K.; KOTZ, S. A.; FRIEDERICI, A. D.; ALTER, K. Pitch modulates lexical identification in spoken word recognition: erp and behavioral evidence. **Cognitive Brain Research**, [S.l.], v.20, p.300–308, July 2004.

FRISCH, S.; HAHNE, A.; FRIEDERICI, A. D. Word category and verb-argument structure information in the dynamics of parsing. **Cognition**, [S.l.], v.91, p.191–219, Apr. 2004.

GALLWITZ, F.; NIEMANN, H.; NöTH, E.; WARNKE, V. Integrated recognition of words and prosodic phrase boundaries. **Speech Communication**, Amsterdam, v.36, p.81–95, Jan. 2002.

GAVAT, I.; ZIRRA, M.; SABAC, B. Pitch Estimation by Block and Instantaneous Methods. **International Journal of Speech Technology**, [S.l.], v.5, p.269–279, 2002.

GODINO-LLORENTE, J. I.; GÓMEZ-VILDA, P. Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors. **IEEE Trans. Biomedical Engineering**, [S.l.], v.51, p.380–384, Feb. 2004.

GOLDSTEIN, H. Formant tracking using the wavelet-based DST. In: IEEE SOUTH AFRICAN SYMPOSIUM COMMUNICATIONS AND SIGNAL PROCESSING, 1994. **Proceedings...** [S.l.: s.n.], 1994. p.183–189.

GOMES, J.; VELHO, L.; GOLDENSTEIN, S. **Wavelets**: teoria, software e aplicações. Rio de Janeiro: IMPA/CNPq, 1997.

GUPTA, M.; GILBERT, A. Speech recognition using artificial neural networks. In: IEEE WORKSHOP AUTOMATIC SPEECH RECOGNITION AND UNDERSTANDING, 2001. **Proceedings...** [S.l.: s.n.], 2001. p.445–448.

GUTSCHALK, A.; PATTERSON, R. D.; SCHERG, M.; UPPENKAMP, S.; RUPP, A. Temporal dynamics of pitch in human auditory cortex. **NeuroImage**, [S.l.], v.22, p.755–766, June 2004.

HAHNE, A.; FRIEDERICI, A. D. Diffential task effects on semantic and syntatic processes as revealed by ERPs. **Cognitive Brain Research**, [S.l.], v.13, p.339–356, 2002.

HASEGAWA-JOHNSON, M.; CHEN, K.; COLE, J.; BORYS, S.; KIM, S.-S.; COHEN, A.; ZHANG, T.; CHOI, J.-Y.; KIM, H.; YOON, T.; CHAVARRIA, S. Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus. **Speech Communication**, Amsterdam, v.46, p.418–439, July 2005.

HASTIE, H. W.; POESIO, M.; ISARD, S. Automatically predicting dialogue structure using prosodic features. **Speech Communication**, Amsterdam, v.36, p.63–79, Jan. 2002.

HAYKIN, S. **Redes Neurais**: princípios e prática. Porto Alegre: Bookman, 2001.

HEIM, S.; OPITZ, B.; FRIEDERICI, A. D. Distributed cortical networks for syntax processing: broca's area as the common denominator. **Brain and Language**, [S.l.], v.85, p.402–408, 2003.

HERRMANN, C. S.; FRIEDERICI, A. D.; ULRICH OERTEL, B. M.; HAHNE, A.; ALTER, K. The brain generates its own sentence melody: a gestalt phenomenon in speech perception. **Brain and Language**, [S.l.], v.85, p.396–401, 2003.

HERTZ, J.; KROGH, A.; PALMER, R. **Introduction to the theory of neural computation**. Reading: Addison-Wesley, 1991.

HIGASHINAKA, R.; SUDOH, K.; NAKANO, M. Incorporating discourse features into confidence scoring of intention recognition results in spoken dialogue systems. **Speech Communication**, Amsterdam, v.48, p.417–436, Mar.-Apr. 2006.

ILLICH, I. Um apelo à pesquisa em cultura escrita leiga. In: OLSON, D. R.; TORRANCE, N. (Ed.). **Cultura Escrita e Oralidade**. São Paulo: Ática, 1995.

JAMES, D. L.; MIIKKULAINEN, R. SARDNET: a self-organizing feature map for sequences. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, NIPS, 7., 1995, Cambridge. **Proceedings...** Cambridge: MIT Press, 1995. v.7, p.577–584.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Upper Saddle River: Prentice Hall, 2000.

KADAMBE, S.; BOUDREAUX-BARTELS, G. A comparison of a wavelet transform event detection pitch detector with classical pitch detectors. In: ASILOMAR CONFERENCE ON SIGNALS, SYSTEMS AND COMPUTERS, 24., 1990. **Proceedings...** [S.l.: s.n.], 1990. v.2, p.1073–1078.

KADAMBE, S.; BOUDREAUX-BARTELS, G. F. Application of the Wavelet Transform for Pitch Detection of Speech Signals. **IEEE Trans. Information Theory**, [S.l.], v.38, p.917–924, 1992.

KADAMBE, S.; SRINIVASAN, P. Adaptive wavelet based phoneme recognition. In: MIDWEST SYMPOSIUM CIRCUITS AND SYSTEMS, 40., 1997. **Proceedings...** [S.l.: s.n.], 1997. v.2, p.720–723.

KAHANE, J. P.; LEMARIÉ-RIEUSSET, P. G. **Fourier booktitle and wavelets**. Sidney: Gordon and Breach, 1995.

KAISER, E.; JOHNSTON, M.; HEEMAN, P. PROFER: predictive, robust finite-state parsing for spoken language. In: IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, ICASSP, 1999. **Proceedings...** [S.l.: s.n.], 1999. v.2, p.629–632.

KIM, K.; YOUN, D. H.; LEE, C. Evaluation of wavelet filters for speech recognition. In: IEEE INT. CONF. SYSTEMS, MAN, AND CYBERNETICS, 2000. **Proceedings...** [S.l.: s.n.], 2000. v.4, p.2891–2894.

KIPP, M. The Neural Path to Dialogue Acts. In: EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE, ECAI, 13., 1998. **Proceedings...** [S.l.]: John Wiley & Sons, 1998. p.175–179.

KOHONEN, T. **Self-Organization and Associative Memory**. [S.l.]: Springer-Verlag, 1984.

KOHONEN, T. The Self-Organizing Map. **Proceedings of the IEEE**, [S.l.], v.78, n.9, p.1464–1480, 1990.

KOMPE, R. **Prosody in Speech Understanding Systems**. Berlin: Springer-Verlag, 1997.

KOTZ, S. A.; MEYER, M.; ALTER, K.; BESSON, M.; CRAMON, D. Y. von; FRIEDERICI, A. D. On the lateralization of emotional prosody: an event-related functional mr investigation. **Brain and Language**, [S.l.], v.86, p.366–376, Sept. 2003.

KURIMO, M. Thematic indexing of spoken documents by using self-organizing maps. **Speech Communication**, Amsterdam, v.38, p.29–45, 2002.

LARSEN, L.; BRØ NDSTED, T.; DYBKJÆR, H.; DYBKJÆR, L.; MUSIC, B.; POVLSEN, C. State-of-the-art of Spoken Language Systems: a survey. In: LARSEN, L. B. (Ed.). **Spoken Dialogue System Reports**. [S.l.]: STC Aalborg University, CCS Roskilde University, CST University of Copenhagen, 1992.

LAVIE, A. GLR*: a robust parser for spontaneously spoken language. In: EUROPEAN SUMMER SCHOOL IN LOGIC, LANGUAGE AND INFORMATION, ESSLLI, 1996, Praga. **Proceedings...** Praga: Carnegie Mellon University, 1996.

LAVIE, A. **GLR***: a robust grammar-focused parser for spontaneously spoken language. 1996. PhD Thesis — Carnegie Mellon University, Pittsburgh.

LETHBRIDGE, T. **Descoberta em macacos estrutura cerebral ligada à fala**. Disponível em: <http://cienciahoje.uol.com.br/controlPanel/materia/view/3662>. Acesso em: ago. 2005.

LEVY, S.; POLLACK, J. Infinite RAAM: a principled connectionist substrate for cognitive modeling. In: INTERNATIONAL CONGRESS OF CHINESE MATHEMATICIANS, ICCM, 2001. **Proceedings...** [S.l.]: Lawrence Erlbaum Associates, 2001.

LEWANDOWSKA-TOMASZCZYK, B. Cross-Linguistic and Language-Specific Aspects of Semantic Prosody. **Language Sciences**, [S.l.], v.18, p.153–178, 1996.

MAYBERRY III, M. R.; MIIKKULAINEN, R. SARDSRN: a neural network shift-reduce parser. In: ANNUAL INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, IJCAI, 16., 1999. **Proceedings...** [S.l.]: Kaufmann, 1999. p.820–825.

MEYER, M.; STEINHAUER, K.; ALTER, K.; FRIEDERICI, A. D.; CRAMON, D. Y. von. Brain activity varies with modulation of dynamic pitch variance in sentence melody. **Brain and Language**, [S.l.], v.89, p.277–289, May 2004.

MINKER, W. Stochastic versus rule-based speech understanding for information retrieval. **Speech Communication**, Amsterdam, v.25, p.223–247, Sept. 1998.

MINKER, W.; GAVALDÀ, M.; WAIBEL, A. Stochastically-based semantic analysis for machine translation. **Computer Speech and Language**, London, v.13, p.177–194, Apr. 1999.

MORETTIN, P. A. **Ondas e ondaletas**: da análise de fourier à analise de ondaletas. Sao Paulo: Edusp, 1999.

MORGAN, N.; BOULARD, H. An Introduction to Hybrid HMM/Connectionist Continuous Speech Recognition. **IEEE Signal Processing Magazine**, [S.l.], p.25–42, May 1995.

MÜLLER, D. N. **Reconhecimento Semântico Através de Redes Neurais Artificiais**. 1996. MSc Thesis — Instituto de Informática, UFRGS, Porto Alegre.

MÜLLER, D. N.; NAVAUX, P. O. A. Representação Computacional da Compreensão da Fala. In: WORKSHOP DE TESES E DISSERTAÇÕES EM INTELIGÊNCIA ARTIFICIAL, WTDIA, 2., 2004. **Anais...** [S.l.]: UFMA, 2004. p.151–160.

MÜLLER, D. N.; NAVAUX, P. O. A. O Uso de Cluster em Sistemas Conexionistas para o Processamento da Compreensão da Fala. In: WORKSHOP DE PROCESSAMENTO PARALELO E DISTRIBUÍDO, WSPPD, 3., 2005. **Anais...** [S.l.]: Instituto de Informática - UFRGS, 2005. p.27–32.

MÜLLER, D. N.; NAVAUX, P. O. A. Sistema para Representação Computacional da Compreensão da Fala. In: CONGRESSO DA SBC - ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, ENIA, 2005. **Anais...** [S.l.]: SBC, 2005. p.1142–1145.

MÜLLER, D. N.; NAVAUX, P. O. A. Computational Model of Speech Understanding. In: IASTED INTERNATIONAL MULTI-CONFERENCE ARTIFICIAL INTELLIGENCE AND APPLICATIONS, AIA, 24., 2006. **Proceedings...** [S.l.]: Acta Press, 2006. p.97–101.

MÜLLER, D. N.; SIQUEIRA, M. L. de; NAVAUX, P. O. A. A Model to Computational Speech Understanding. In: COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE: INTERNATIONAL WORKSHOP, PROPOR, 7., 2006, Berlin. **Proceedings...** Berlin: Springer, 2006. p.216 – 219.

MÜLLER, D. N.; SIQUEIRA, M. L. de; NAVAUX, P. O. A. **A Connectionist Approach to Speech Understanding**. Aceito para apresentação na International Joint Conference on Neural Networks, IJCNN, 2006.

NEAL, J. G.; SHAPIRO, S. C. Knowledge-Based Parsing. In: BOLC, L. (Ed.). **Natural Language Parsing Systems**. Berlim: Springer-Verlag, 1987.

NÖTH, E.; BATLINER, A.; WARNKE, V.; HAAS, J.; BOROS, M.; BUCKOW, J.; HUBER, R.; GALLWITZ, F.; NUTT, M.; NIEMANN, H. On the use of prosody in automatic dialogue understanding. **Speech Communication**, Amsterdam, v.36, p.45–62, Jan. 2002.

OBAIDAT, M. S.; LEE, C.; SADOUN, B.; NELSON, D. Estimation of pitch period of speech signal using a new dyadic wavelet algorithm. **Information Sciences**, [S.l.], v.119, p.21–39, 1999.

OLIVEIRA FILHO, K. d. S. **Fundamentos de Radiodiagnóstico por Imagem**. Disponível em: <http://www.if.ufrgs.br/ast/med/imagens>. Acesso em: ago. 2005.

OLIVEIRA, P. M. T. **Auxílio Visual À Oralização de Surdos**. 1998. MSc Thesis — UFRJ, Rio de Janeiro.

PACHECO, H. C. F.; DILLINGER, M.; CARVALHO, M. L. de. Uma nova abordagem para a análise sintática do português. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DE PORTUGUÊS ESCRITO E FALADO, 2., 1996, Curitiba. **Anais...** Curitiba: CEFET-PR, 1996. p.51–60.

PARGELLIS, A.; FOSLER-LUSSIER, E.; LEE, C.-H.; POTAMIANOS, A.; TSAI, A. Auto-induced semantic classes. **Speech Communication**, Amsterdam, v.43, p.183–203, Aug. 2004.

PELL, M. D. Cerebral mechanisms for understanding emotional prosody in speech. **Brain and Language**, [S.l.], p.221–234, Feb. 2006.

PITZ, M.; NEY, H. Vocal Tract Normalization Equals Linear Transformation in Cepstral Space. **IEEE Trans. Speech Audio Processing**, [S.l.], v.13, p.930–944, Sept. 2005.

POLLACK, J. B. Recursive Distributed Representations. **Artificial Intelligence**, [S.l.], v.33, p.77–105, 1990.

RABINER, L. R.; JUANG, B. H. **Fundamentals of Speech Recognition**. Englewood Cliffs: Prentice Hall, 1993.

RABINER, L.; SCHAFER, R. W. **Digital Processing of Speech Signals**. Upper Saddle River: Prentice-Hall, 1978.

RAYMOND, C.; BÉCHET, F.; MORI, R. D.; DAMNATI, G. On the use of finite state transducers for semantic interpretation. **Speech Communication**, Amsterdam, v.48, p.288–304, Mar. 2006.

RICOTTI, L. Multitapering and a wavelet variant of MFCC in speech recognition. **IEEE Trans. Vision, Image and Signal Processing**, [S.l.], v.152, p.29–35, Feb. 2005.

RITTER, H.; MARTINEZ, T.; SCHULTEN, K. **Neural Computation and Self-Organizing Maps**. New York: Addison-Wesley, 1992.

ROSSET, S.; PASTORELLO, B.; JUNIOR, D. A.; CARNEIRO, A. A. **Pesquisa de métodos não invasivos para mapear regiões do cérebro pode ser usada em cirurgias de pacientes com epilepsia**. Disponível em: <http://www.canalciencia.ibict.br/pesquisas/pesquisa.php?ref_pesquisa=134>. Acesso em: nov. 2005.

ROSSI, S.; GUGLER, M. F.; HAHNE, A.; FRIEDERICI, A. D. When word category information encounters morphosyntax: an erp study. **Neuroscience Letters**, [S.l.], v.384, p.228–233, Aug. 2005.

SAYOOD, K. **Introduction to data compression**. San Francisco: Morgan Kaufmann, 1996.

SCHIRMER, A.; KOTZ, S. A.; FRIEDERICI, A. D. Sex differentiates the role of emotional prosody during word processing. **Cognitive Brain Research**, [S.l.], v.14, p.228–233, Aug. 2002.

SCHIRMER, A.; KOTZ, S. A.; FRIEDERICI, A. D. On the role of attention for the processing of emotions in speech: sex differences revisited. **Cognitive Brain Research**, [S.l.], v.24, p.442–452, Aug. 2005.

SCHIRMER, A.; ZYSSET, S.; KOTZ, S. A.; CRAMON, D. Y. von. Gender differences in the activation of inferior frontal cortex during emotional speech perception. **NeuroImage**, [S.l.], v.21, p.1114–1123, Mar. 2004.

SCOTT, S. K.; WISE, R. J. S. Functional imaging and language: a critical guide to methodology and analysis. **Speech Communication**, Amsterdam, v.41, p.7–21, 2003.

SHAMMA, S. A.; MORRISH, K. A. Synchrony suppression in complex stimulus responses of a biophysical model of the cochlea. **J. Acoustic Soc. American**, [S.l.], v.81, p.1486–1498, May 1987.

SIKKEL, K. **A Framework for Specification and Analysis of Parsing Algorithms**. Berlim: Springer-Verlag, 1997.

SMITH, S. W. **The scientist and engineer's guide to digital speech signal processing**. San Diego: California Technical Publishing, 1999.

STEEDMAN, M. Connectionist Sentence Processing in Perspective. **Cognitive Science**, [S.l.], v.23, p.615–634, 1999.

STOLCKE, A.; SHRIBERG, E.; BATES, R.; COCCARO, N.; JURAFSKY, D.; MARTIN, R.; METEER, M.; RIES, K.; TAYLOR, P.; ESS-DYKEMA, C. V. Dialog Act Modeling for Conversational Speech. In: SPRING SYMPOSIUM ON APPLYING MACHINE LEARNING TO DISCOURSE PROCESSING, AAAI, 1998, Menlo Park. **Proceedings...** Menlo Park: AAAI Press, 1998. p.98–105.

STOWE, L. A.; PAANS, A. M. J.; WIJERS, A. A.; ZWARTS, F. Activations of 'motor' and other non-language structures during sentence comprehension. **Brain and Language**, [S.l.], v.89, p.290–299, May 2004.

TEBELSKIS, J. **Speech Recognition using Neural Networks**. 1995. PhD Thesis — Carnegie Mellon University, Pittsburgh.

TRAN, T.; HA, Q.; DISSANAYAKE, G. New Wavelet-Based Pitch Detection Method for Human-Robot Voice Interface. In: IEEE/RSJ INTERNATIONAL CONFERENCE ON INTELLIGENT ROBOTS AND SYSTEMS, IROS, 2004. **Proceedings...** [S.l.: s.n.], 2004. v.1, p.527–532.

TUFEKCI, Z.; GOWDY, J. Feature extraction using discrete wavelet transform for speech recognition. In: IEEE SOUTHEASTCON, 2000. **Proceedings...** [S.l.: s.n.], 2000. p.116–123.

VALIATI, J. F. **Reconhecimento de voz para comandos de direcionamento por meio de redes neurais**. 2000. MSc Thesis — Instituto de Informática, UFRGS, Porto Alegre.

VOS, S. H.; FRIEDERICI, A. D. Intersentential syntactic context effects on comprehension: the role of working memory. **Cognitive Brain Research**, [S.l.], v.16, p.111–122, 2003.

WAMBACQA, I. J. A.; JERGER, J. F. Processing of affective prosody and lexical-semantics in spoken utterances as differentiated by event-related potentials. **Cognitive Brain Research**, [S.l.], v.20, p.427–437, 2004.

WARD, N. Pragmatic Functions of Prosodic Features in Non-Lexical Utterances. In: SPEECH PROSODY, 2004. **Proceedings...** [S.l.]: ISCA Archive, 2004.

WEBER, V.; WERMTER, S. Using hybrid connectionist learning for speech/language analysis. In: WERMTER, S.; RILOFF, E.; SCHELER, G. (Ed.). **Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing**. [S.l.]: Springer Verlag, 1996. p.87–101.

WERMTER, S.; LÖCHEL, M. Learning Dialog Act Processing. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 1996, Kopenhagen. **Proceedings...** Kopenhagen: [s.n.], 1996. p.740–745.

WERMTER, S.; WEBER, V. SCREEN: learning a flat syntactic and semantic spoken language analysis using artificial neural networks. **Journal of Artificial Intelligence Research**, [S.l.], v.6, n.1, p.35–85, 1997.

WITTMANN, L. H.; RIBEIRO, R. D. Recursos Lingüísticos e Processamento Morfológico do Português: o palavroso e o projecto le-parole. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DE PORTUGUÊS ESCRITO E FALADO, 2., 1998, Porto Alegre. **Anais...** Porto Alegre: Todeschini, 1998.

WRIGHT, H. F. **Modelling Prosodic and Dialogue Information for Automatic Speech Recognition**. 1999. PhD Thesis — University of Edimburg, Edimburg.

YANG, X.; WANG, K.; SHAMMA, S. Auditory representations of acoustic signals. **IEEE Trans. Information Theory**, [S.l.], v.38, p.824–839, Mar. 1992.

YAO, J.; ZHANG, Y.-T. The application of bionic wavelet transform to speech signal processing in cochlear implants using neural network simulations. **IEEE Trans. Biomedical Engineering**, [S.l.], v.49, p.1299–1309, Nov. 2002.

ZECHNER, K.; WAIBEL, A. Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 36., AND INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 17., COLING/ACL, 1998, Montreal. **Proceedings...** Montreal: [s.n.], 1998.

ZHANG, T.; HASEGAWA-JOHNSON, M.; LEVINSON, S. E. Extraction of pragmatic and semantic salience from spontaneous spoken English. **Speech Communication**, Amsterdam, v.48, p.437–462, Mar.-Apr. 2006.

ANEXO A GERAÇÃO DE COEFICIENTES FONÉTICOS

```
%
% Copyright@2005 Daniel Nehme Müller
%
% COMFALA – Modelo computacional para compreensão da fala
%
% Tese de doutorado – PPGC – II - UFRGS
%
% Subsistema matlab para geração de coeficientes fonéticos
%


% le arquivo
nomearq = input('arquivo: ','s');
[voz,Fs,bits]  = wavread(nomearq);


% decomposicao ondeletas
nivel = 6;
filtro = 'db4';
entropia = 'shannon';
quant = 2^nivel;
[t,d] = wpdec(voz,nivel,filtro,entropia);
for i = 0:quant-1
      faixa =  [nivel i];
      cfs = wdatamgr('read_cfs',d,t,faixa);

    tamcfs = length(cfs);
      soma = sum(cfs.^2); % soma dos quadrados - energia

    energia(i+1) = log(soma/tamcfs)/10;
end;
menor = min(energia);
energia = energia + abs(menor);


% grava coeficientes
nomearq2 = sprintf('%s_cw',nomearq);
arq = fopen(nomearq2,'w');
fprintf(arq,'%s\n',nomearq);
for i = 1:quant

   fprintf(arq,'%f ',energia(i));
end
fprintf(arq,'\n');
fclose(arq);
```

ANEXO B GERAÇÃO DE COEFICIENTES PROSÓDICOS

```
%
% Copyright@2005 Daniel Nehme Müller
%
% COMFALA – Modelo computacional para compreensão da fala
%
% Tese de doutorado – PPGC – II - UFRGS
%
% Subsistema matlab para geração de coeficientes prosódicos
%


nomearq = input('arquivo: ','s');
[voz,Fs,bits]  = wavread(nomearq);
tam = length(voz);


% decomposição ondeletas
nivel = 2;
filtro = 'db4';
cca = voz;
for i = 1:nivel

    [cca,ccd] = dwt(cca,filtro);
end


[cca2,ccd2] = dwt(cca,filtro);  % pega derivada - proxima escala


ccadec = dyaddown(cca);  % decima para ficar compatível com cca2


tamcca = length(ccadec);


% pega os pontos da derivada maiores que 80% da derivadora
contasaida = 0;
for i = 1:tamcca

   if abs(cca2(i))>=abs(ccadec(i)*.8)

      contasaida = contasaida +1;

      saida(contasaida)=ccadec(i);

   end
end


tamsaida = length(saida);
pontos = zeros(tamsaida,1);
pontosf = pontos;
freq = pontos;


if saida(1) >= 0   % inicializa verificacao de ciclos

    atual = 0;
```

```
  if atual ~= anterior
      pontosf(i) = saida(i);
      anterior = atual;
      troca = troca+1;
   end

   if troca == 3
      pontos(i) = saida(i);
      altmarca = i - marca;
      freq(i) = 1/altmarca;
      marca = i;
      troca = 1;
   end
end

% retira pontos zerados
x = 1;
for i = 1:tamsaida
    if pontos(i) ~= 0
       pnovo(x) = pontos(i);
       x=x+1;
    end
end

% retira frequencias zeradas
for i = 2:tamsaida
    if freq(i) == 0
       freq(i) = freq(i-1);
    end
end
```