Sistema para Representação Computacional da Compreensão da Fala

Daniel Nehme Müller, Philippe O. A. Navaux[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul

{danielnm,navaux}@inf.ufrgs.br

Abstract. *This paper proposes a computational model to speech comprehension based on neurocognitive researches. The computational representation uses several techniques as wavelets transform and connectionists models. The speech signal codification and data prosodic extraction are derived from wavelets coefficients. The connectionist models are used to perform syntactic parsing and prosodic-semantic mapping.*

Resumo. *O presente trabalho apresenta um modelo computacional para compreensão da fala baseado em estudos neurocognitivos. Para uma adequada representação computacional foram utilizadas várias técnicas como a transformada de ondeletas (wavelets transform) e modelos conexionistas. Através das ondeletas realiza-se a codificação do sinal de fala e extração de dados prosódicos. Os modelos conexionistas são usados para o parsing sintático e mapeamento prosódico-semântico.*

1. Compreensão de Fala

O presente trabalho sustenta que é possível integrar diversos subsistemas para representar todo o processo de compreensão da fala, desde a captura do sinal até a decisão do contexto semântico da fala. Baseando-se em modelo biológico (ver seção 2.), procura-se verificar se é possível uma codificação temporal do sinal de fala, o que auxilia na definição do contexto prosódico do que é falado. Deseja-se ainda apresentar uma solução conexionista para o parsing lingüístico, através do uso da codificação da fala para definir a estrutura das construções da linguagem. Para as representações semântica e prosódica também optou-se por usar redes neurais artificiais, o que permite a criação de mapas conceituais que auxiliarão na definição dos contextos da fala. Para o processamento do sinal de fala e segmentação prosódica, foi escolhida a técnica de transformada ondeletas (*wavelets transform*) descrita na seção 3. Os demais subsistemas conexionistas usados no parsing sintático e definição de contextos semânticos estão descritos na seção 4. Na seção 5. são apresentados alguns resultados do presente trabalho.

2. O Modelo Biológico

O sistema aqui apresentado parte do princípio que um estudo adequado dos processos biológicos envolvidos na compreensão de fala deverá levar a uma coerente e válida representação computacional de tais processos. Para tanto, buscou-se uma base biológica nos recentes estudos por visualização de ressonância magnética funcional (fMRI) e tomografia por visualização de pósitrons (PET). Estes estudos foram encontrados no trabalho

pioneiro dos pesquisadores Angela Friederici e Kai Alter, do Instituto Max Planck de Neurociência Cognitiva. Friederici [Friederici 2002] propôs um *Modelo Neurocognitivo de Processamento de Frases Ouvidas*, que é o princípio do sistema aqui descrito. A partir deste modelo neurocognitivo, este trabalho propõe diversas técnicas computacionais que possibilitem uma adequada representação do processo, descritas na seção 4. Desta forma, Friederici mapeou quais partes do cérebro eram ativadas no tempo, dados os diferentes testes aplicados. Ela dividiu o processamento das frases ouvidas em 4 grandes fases: segmentação fonológica e seqüenciamento (até 100 ms); construção da estrutura sintática (entre 100 e 300 ms); definição das relações semânticas e foco temático (entre 300 e 500 ms); integração sintática-semântica-prosódica (entre 500 e 1000 ms) [Friederici 2002] [Friederici and Alter 2004].

3. Ondeletas

A codificação tradicional de elementos da fala utiliza a transformada de Fourier para representação das características da fala. No presente trabalho é usada a transformada ondeletas, também conhecida como *wavelets*. A análise de multiresolução em ondeletas foi utilizada neste trabalho para extração das características do sinal de fala. Estas características foram divididas em lingüísticas e prosódicas. As lingüísticas são obtidas a partir de coeficientes de ondeletas da decomposição do sinal da fala. As prosódicas são extraídas a partir da variação da onda fundamental. Para se obter informações sobre as variações da onda fundamental da fala, necessita-se detectar os pontos de máximo das ondeletas, os quais correspondem a instantes de fechamento da glote (GCI) [Kadambe and Boudreaux-Bartels 1990]. Uma vez obtidos os pontos de máximo, tem-se a identificação da onda fundamental. Além da codificação lingüística, através da onda fundamental é possível a identificação de características prosódicas. Estas características são identificadas pela análise do comportamento da fundamental, que será aplicada aos mapas semântico e prosódico.

4. O Modelo Computacional

O modelo computacional proposto utiliza técnicas de ondeletas e conexionistas que buscam atender às necessidades de representação do modelo biológico descrito. A figura 1 representa o modelo computacional com suas quatro fases. Na primeira fase do modelo computacional, o sinal é processado pela aplicação de transformadas de ondeletas. Os coeficientes resultantes deste processamento dão informações sobre o comportamento do pitch e são usados como entrada nas redes neurais das fases seguintes. Este processamento será detalhado na seção 3. A segunda fase computacional é a aplicação dos coeficientes de ondeletas para a geração de registros temporais e árvores de parsing através do sistema SARDSRN-RAAM [Mayberry III and Miikkulainen 1999]. Este sistema proposto é baseado no SARDSRN-RAAM, que é formado pelas redes SRN (*Simple Recurrent Network*) [Elman 1990], SARDNET (*Sequential Activation Retention and Decay Network*) [James and Miikkulainen 1995] e RAAM (*Recursive Auto-Associative Memory*) [Pollack 1990]. Na terceira fase ocorre a criação de mapas semântico e prosódico utilizando-se os Mapas de Características Auto-Organizáveis (SOM - *Self-Organizing Map*) [Kohonen 1984], que permitem o agrupamento de padrões com características similares. A quarta fase computacional realiza a recepção e análise das saídas das segunda e terceira fases. Nessa fase o modelo indica, para dada frase interpretada, qual o contexto semântico mais provável.

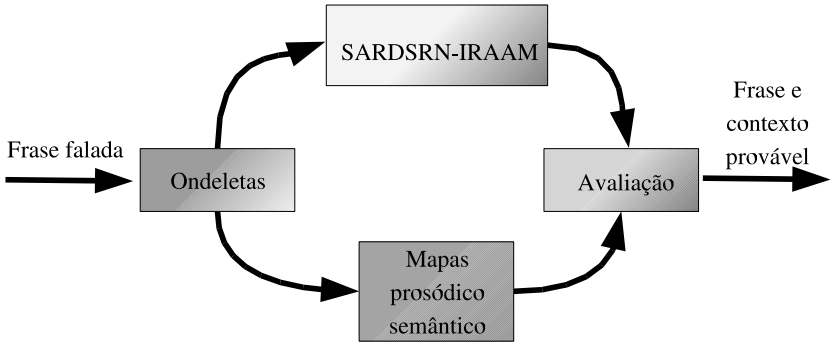

Figura 1. Modelo computacional proposto.

O processamento do parsing inicia com o uso da codificação lingüística obtida da transformada ondeletas. As palavras codificadas são estruturadas através da RAAM, cuja ativação permite o seqüenciamento das palavras na frase pelo SARDSRN-RAAM e por fim gera a representação da análise parcial, que é decodificada, novamente pela rede RAAM. A entrada de treinamento deste modelo são seqüências de palavras e suas saídas esperadas. A entrada de testes são seqüências de palavras, nem sempre com nexo e com organização diferente do treinamento. A maior característica do SARDSRN-RAAM é sua grande capacidade de geração de seqüências de parsing, que permitirão o treinamento (e posterior reconhecimento) de múltiplas árvores de parser. Estas árvores estão comprimidas na rede RAAM, o que permite uma análise (reconhecimento) muito mais rápida e robusta que os métodos simbólicos. O processamento prosódico-semântico compõe-se basicamente de redes SOM: uma para o mapeamento das características prosódicas das palavras; outra para o aprendizado das relações semânticas das palavras; outra para a identificação dos agrupamentos -semânticos das palavras; e uma última para a organização de grupos de frases dentro dos mesmos contextos prosódico-semânticos. O posterior reconhecimento é feito a partir do último mapa, referente às relações semânticas de frases, o qual indica a frase-padrão (com um significado definido) que é semelhante à frase a identificar. Com este mecanismo básico é possível realizar-se o *reconhecimento do significado* de quaisquer frases dentro do contexto pré-definido, sem a necessidade de formalização e definição de regras lingüísticas ou suas relações, como ocorre em muitos sistemas convencionais para reconhecimento de língua natural [Müller 1996]. O mapa prosódico agrupa as palavras segundo os sinais derivados da análise de variações na onda fundamental, que dão informações sobre as pausas e acentos da linguagem. O mapa semântico de palavras organiza grupos de palavras segundo sua estrutura lingüística e sua pronúncia. O mapa semântico-prosódico utiliza a indicação dos neurônios ativados no mapa semântico, mais a codificação de contexto indicada pelo neurônio ativado no mapa prosódico, para formar grupos de contextos de palavras. O mapa de frases realiza agrupamentos segundo o conjunto de palavras mais prováveis dentro de um contexto de expressão.

O sistema que simula o modelo inicia com a captura do sinal de fala, o qual é armazenado em arquivo no formato *wave*. Este sinal é processado pelo subsistema que realiza a transformada ondeletas e gera os coeficientes lingüísticos e os prosódicos. O subsistema SARDSRN-RAAM é treinado a partir dos coeficientes lingüísticos, gerando a temporização e a árvore de parsing. Isto permite a geração de estruturas frasais a partir

da entrada apresentada. Para o reconhecimento, é apresentada uma seqüência de códigos e o subsistema apresentará a frase mais provável. O subsistema dos mapas prosódicos recebe os coeficientes prosódicos e lingüísticos e gera agrupamentos como explicados na seção anterior. Na etapa de reconhecimento, a seqüência de códigos apresentados aos mapas indicam na saída o contexto mais provável da frase. Analisando-se a saída do SARDSRN-RAAM e dos mapas, tem-se, por um lado, hipóteses de construções de frases e, de outro, seu contexto mais provável. Unindo ambos, pode-se escolher qual a melhor frase treinada ou contexto se adequa à frase apresentada para reconhecimento.

5. Resultados

O trabalho aqui apresentado, que vem sendo desenvolvido como tese de doutorado na Universidade Federal do Rio Grande do Sul, utiliza-se da inspiração biológica e da combinação de modelos conexionistas para propor um modelo de análise computacional da compreensão de fala. Como exemplo de seu funcionamento, pode-se citar frases *a menina mordeu o cachorro* e *o gato perseguiu a menina* que não estão no conjunto de treinamento. A primeira frase apresentou a aceitação, com erro 0,34 o subsistema de sintaxe. Já no mapa de frases, obteve um posicionamento eqüidistante com os verbos *gostou* e *mordeu*. Já a segunda frase, contendo o verbo perseguiu, obteve rejeição no reconhecimento sintático com erro 0,89, mas no mapa de frases mostrou-se próximo ao verbo *mordeu*. Estes dois exemplos significam que, no caso da primeira frase, que ela correspondeu a outras frases sintaticamente corretas com o verbo *morder*, o que facilita a decisão no campo semântico, que ficou entre *gostar* e *morder*. No caso da segunda frase, o sistema aproxima apenas para o contexto das frases com o verbo *morder*, compensando a margem de erro do reconhecimento sintático.

Referências

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, 6:78–84.

Friederici, A. D. and Alter, K. (2004). Lateralization of auditory language functions: A dynamic dual pathway model. *Brain and Language*, 89:267–276.

James, D. L. and Miikkulainen, R. (1995). SARDNET: A self-organizing feature map for sequences. In Tesauro, G. e. a., editor, *Advances in Neural Information Processing Systems 7 (NIPS'94, Denver, CO)*, pages 577–584. MIT Press.

Kadambe, S. and Boudreaux-Bartels, G. (1990). A comparison of a wavelet transform event detection pitch detector with classical pitch detectors. *Twenty-Fourth Asilomar Conference on Signals, Systems and Computers*, 2:1073–1078.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Springer-Verlag.

Mayberry III, M. R. and Miikkulainen, R. (1999). SARDSRN: a neural network shift-reduce parser. In *Proceedings of IJCAI-99*, pages 820–825. Kaufmann.

Müller, D. N. (1996). Reconhecimento Semântico Através de Redes Neurais Artificiais. Master's thesis, CPGCC-UFRGS.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, 33:77–105.